

From Far-Field Dynamics to Close-Up Confidence: Action Recognition Across Varying Camera Distances

by

Kseniia Buzko

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Systems Design Engineering

Waterloo, Ontario, Canada, 2025

© Kseniia Buzko 2025

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

This thesis incorporates material from one paper for which I was the lead author. I was responsible for the conceptualization, code development, manuscript preparation, and submission.

Chapter 3 contains material from the following previously published work:

Buzko, K., Nazemi, A., Clausi, D., & Chen, Y. (2025). *Ice Hockey Action Recognition via Contextual Priors. Linköping Hockey Analytics Conference (LINHAC)*. doi:[10.3384/ecp214001](https://doi.org/10.3384/ecp214001).

Abstract

Human action recognition (HAR) refers to the task of identifying and classifying human actions within videos or sequences of images. This field has gained significant importance due to its diverse applicability across various domains, such as sports analytics, human-computer interaction, surveillance, and interpersonal communication. Accurate action recognition becomes especially difficult when the camera distance changes, because the cues that matter shift with scale. For instance, a close-up hinges on facial emotion (such as smiles and eye gaze), whereas a medium shot relies on hand gestures or objects being manipulated.

In the context of **HAR**, we distinguish two primary scenarios that illustrate this challenge. The first is the **far-field scenario** setting, characterized by subjects positioned at a distance and often exhibiting rapid movement, which leads to frequent occlusions. This scenario is commonly observed in sports broadcasts, where capturing the game’s dynamics is essential. In contrast, the **near-field scenario** setting involves subjects that are nearby and tend to remain relatively static. This setting enables the capture of subtle yet informative gestures, similar to those observed in presenter-focused videos. Although most studies treat these regimes separately, modern media (films, replays, vlogs) cut or zoom fluidly between them. An effective recognizer must therefore decide dynamically which cues to prioritize: facial emotion in tight close-ups, hand or torso motion in medium shots, and full-body dynamics in wide views.

Despite substantial progress, current **HAR** pipelines rarely adapt across that zoom continuum. This thesis, therefore, asks: *What scale-specific hurdles confront human action recognition in **far-field scenario**, **near-field scenario**, and **zoom-mixed scenarios**, and how can insights from separate case studies keep recognition robust when the camera sweeps from full-body scenes to tight close-ups and back again?*

To answer, we contribute three scale-aware systems:

1. **Hockey Action Identification and Keypose Understanding (HAIKYU) (far-field scenario)**. For hockey broadcasts, we introduce temporal **bounding-box normalization**, which removes camera-induced scale jitter, and a 15-**keypoints** skeleton that adds stick endpoints. Combined with normalization, this improves **Top-1** accuracy from 31% to 64%, showing that stick cues are indispensable for ice-hockey actions.
2. **Confidence Fostering Identity-preserving DyNamic Transformer (CONFIDANT) (near-field scenario)**. We curate a 38-class **micro-gestures** dataset

and train an upper-body action recognizer that flags unconfident cues, such as folding arms, crossing fingers, and clasping hands. A diffusion-based video editor then rewrites these segments into confident counterparts, serving as a downstream demonstration of fine-grained recognition.

3. **Scale-aware routing framework for mixed-zoom action recognition (Zoom-Gate) (zoom-mixed scenario)**. A lightweight zoom score derived from the bounding box area and the density of detected [keypoints](#) routes each [tracklet](#) to the specialist model best suited to that scale. Experiments confirm that this scale-aware routing, combined with context-specific skeletons, delivers robust performance across mixed-zoom datasets.

Collectively, these contributions demonstrate that coupling scale-aware preprocessing with context-specific skeletons can maintain pose-centric [HAR](#) reliability across the zoom spectrum. The resulting frameworks open avenues for real-time segmentation, multiview fusion, and ultimately a unified, scale-invariant action understanding pipeline.

Acknowledgements

First and foremost, I would like to express my sincerest gratitude to my supervisors. Professor David, your mentorship, kindness, and guidance made my journey possible. Professor Yuhao, your inspiration and fresh perspectives have significantly shaped my development, and I'm thankful for your trust and motivation. I am also grateful to be part of the Vision and Image Processing (VIP) Research Group, which has provided both intellectual stimulation and personal encouragement.

I owe infinite gratitude to my family, whose love has been my anchor. Mom, words can never fully capture how grateful I am to be your daughter. Your unwavering faith in me, even in moments when I doubted myself, has been my greatest source of strength. Your sacrifices and unconditional love have given me the courage to keep moving forward, and I love you more than I can ever express. Grandma, I am endlessly thankful for your wisdom, your gentle advice, and your ever-present willingness to listen. The way you raised me has shaped me into a better person, and I carry the lessons you instilled in me every day. For this, for being part of this family, I will always remain profoundly grateful.

To my dearest friends of many years, Masha, Misha, and Egor, thank you for being my chosen family for more than a decade. Distance has never weakened our bond. Even though we now live in different cities and countries, you have always believed in me and supported each step I have taken. I eagerly await the day we gather again to laugh, joke, and play Alias together like old times. Masha, you are my best friend for life, and I am deeply grateful for the rare and beautiful gift of our friendship, a bond that truly happens only once in a lifetime.

To my new friends in Canada, Chang, Olha, and Marta, I am so grateful for the gift of your friendship. Thank you for your kindness, for listening to me on my hardest days, and for sharing in moments of joy. You are all such remarkable people, and I know that each of you will achieve amazing things. I feel truly fortunate to have found you here.

Finally, with all my heart, I thank the Ukrainian army. Without your courage and sacrifice, I would not be here today writing these words. You protect not only my family and loved ones but also the very possibility of a future filled with peace, hope, and opportunity. My gratitude to you is beyond measure.

Борітеся — Поборете!
Keep fighting — you will prevail!

Dedication

I dedicate this thesis to my grandfather, Vasyl Karpushenko. You will always remain the person I look up to, both in life and in science. Your journey, marked by talent, intelligence, and determination, has left an everlasting impression on me and continues to inspire me every day. You rose to become an admirable and respected figure in your field, and I can only hope that one day I may live up to the example you set. More than anything, I hope that my achievements make you proud, just as I have always been proud to be your granddaughter.

I still remember the phrase you once taught me that has since become my motto in life, sport, and science. As a child, I hated picking up the rotting apples in the yard, and I complained about how unpleasant it was. You simply told me: "Time is relative—fifteen minutes of work, and then you can wash your hands and move on." These words have stayed with me ever since. Now, whenever I face challenges, I remind myself that hardships pass quickly, and on the other side there is rest and renewal.

With all my love and gratitude, I dedicate this work to you, Grandpa, may it honor your memory and make you proud.

Table of Contents

Author’s Declaration	ii
Statement of Contributions	iii
Abstract	iv
Acknowledgements	vi
Dedication	vii
List of Figures	xiii
List of Tables	xv
List of Abbreviations	xvi
List of Symbols	xviii
1 Introduction	1
1.1 Motivation	2
1.2 Challenges	4
1.2.1 Far-field scenarios	4
1.2.2 Near-field scenarios	4

1.2.3	Zoom-mixed scenarios	5
1.3	Research Objectives	5
1.4	Contributions	6
1.5	Outline of the Thesis	6
2	Background Research	8
2.1	Human Pose Estimation and Skeleton-Based Action Recognition	8
2.2	Far-field scenario Action Recognition	10
2.3	Near-field scenario Action Recognition	12
2.4	Zoom-mixed scenario Action Recognition	13
2.5	Contextual and Object-Aware Action Recognition	14
2.6	Generative Video Editing	16
2.7	Conclusion	17
3	Methodology (Far-Field Scenario: HAIKYU)	18
3.1	Ice Hockey Dataset	19
3.2	Methodology	20
3.2.1	Pipeline Overview	20
3.2.2	Preprocessing	21
3.2.3	Bounding Box Normalization	21
3.2.4	Pose Keypoint Extraction	22
3.2.5	GCN for Action Recognition	24
3.3	Experiments	24
3.3.1	Experimental Setup	24
3.3.2	Experimental Results	24
3.3.3	Qualitative Results	25
3.4	Conclusion	28

4	Methodology (Near-Field Scenario: CONFIDANT)	29
4.1	CONFIDANT Dataset	29
4.1.1	Existing Micro-Gesture Datasets	29
4.1.2	CONFIDANT Dataset Development	30
4.1.3	Classification into Confident vs Unconfident	31
4.2	Micro-Gesture Action Recognition Model	32
4.2.1	Problem Formulation	33
4.2.2	Pose Representation and Keypoint Design	33
4.2.3	Model Architecture	34
4.2.4	Experimental Setup	34
4.2.5	Experimental Evaluation	34
4.3	Application: Identity-Preserving Confidence Editing (Downstream Task) .	37
4.3.1	Preprocessing	38
4.3.2	Diffusion Model for Confident Gesture Editing	38
4.3.3	Diffusion Model for Facial Expression and Lip Movement Generation	39
4.3.4	Experimental Setup	40
4.3.5	Quantitative Results	40
4.3.6	Qualitative Results	41
4.3.7	Ethical Statement and Risks of Misuse	41
4.4	Conclusion	42
5	ZoomGate: Scale-Aware Action Recognition Across Mixed Zoom Levels	44
5.1	Problem Definition & Motivation	45
5.2	Semi-Supervised Movie Dataset	46
5.2.1	Semi-Supervised Annotation Process	46
5.2.2	Dataset Statistics	46
5.2.3	Manual Dataset Verification	47
5.3	Pipeline Overview	47

5.3.1	Stage 1: Person Detection and Tracking	48
5.3.2	Stage 2: Keypoint Extraction	48
5.3.3	Stage 3: Heuristic View Classification	49
5.3.4	Stage 4: Temporal Post-Processing	49
5.3.5	Stage 5: Action Recognition Routing	49
5.4	Heuristic Rules for View Classification	50
5.4.1	Baseline Rules	50
5.4.2	Failure Cases and Edge Conditions	51
5.4.3	Updated Rules	51
5.4.4	Temporal Smoothing	53
5.4.5	Qualitative Results for Zoom Labels	54
5.5	Action Recognition	54
5.5.1	Dataset and Statistics	55
5.5.2	Hand-Only Gesture Recognition	55
5.5.3	Facial Emotion Recognition	57
5.5.4	Upper-body Action Recognition	58
5.5.5	Full-body Action Recognition	59
5.6	Conclusion	59
6	Conclusion and Future Work	63
6.1	Summary of Contributions	63
6.2	Future Work	64
6.2.1	Far-Field Scenario Improvements	64
6.2.2	Near-Field Scenario Enhancements	65
6.2.3	Mixed-Scale and Zoom-Aware Extensions	65
6.2.4	Toward Unified Action Recognition	66
	References	67

APPENDICES	73
A Dataset Card: HAIKYU	74
B ChatGPT Classification for CONFIDANT Dataset	75
C Dataset Card: CONFIDANT	77
D Resource Note: ZoomGate	78
E Pseudocode Listings	79
F Hand Gesture Label Set	81
Glossary	82

List of Figures

1.1	Near-field vs. Far-field Scenarios	3
2.1	COCO vs. COCO-WholeBody skeletons	9
2.2	GCN-based recognition pipeline	10
2.3	Bounding-box sizes in far-field sports	11
2.4	Near-field micro-gesture examples	12
2.5	MovieNet shot-scale taxonomy	15
3.1	Contextual priors in HAIKYU	19
3.2	HAIKYU pipeline overview	21
3.3	Effect of bounding-box normalization	23
3.4	Skeletal configurations in HAIKYU	23
3.5	HAIKYU confusion matrices	26
3.6	HAIKYU qualitative predictions	27
4.1	CONFIDANT dataset gesture examples	32
4.2	CONFIDANT qualitative predictions	36
4.3	CONFIDANT generative editing pipeline	37
4.4	Frame connection process	38
4.5	Diffusion models for gesture editing	39
4.6	Qualitative editing results in CONFIDANT	42

5.1	GUI for manual verification	47
5.2	ZoomGate pipeline overview	48
5.3	Examples of five zoom-view categories	50
5.4	Failure cases in zoom classification	52
5.5	Qualitative zoom-view labeling results	55
5.6	Qualitative examples of HANDSONLY predictions	57
5.7	Facial emotion recognition results	58
5.8	Upper-body recognition results	60
5.9	Full-body recognition results	61

List of Tables

3.1	Ice-hockey action dataset classes	20
3.2	HAIKYU accuracy across skeletons	25
4.1	CONFIDANT dataset statistics	31
4.2	Training hyperparameters for CONFIDANT	34
4.3	CONFIDANT recognition accuracies	35
4.4	CONFIDANT quantitative evaluation	41
5.1	Semi-supervised trailer dataset statistics	46
5.2	Baseline heuristic rules	51
5.3	Updated heuristic rules	53
5.4	Action recognition dataset statistics	56
B.1	Classification of Micro-Gestures Using ChatGPT	76
F.1	Adapted hand gesture labels for cinematic scenarios.	81

List of Abbreviations

2s-AGCN two-stream adaptive graph convolutional network 9, 24

BB bounding box 6, 7, 19, 21, 22, 46–48, 52, 53, 64, 65

CONFIDANT Confidence Fostering Identity-preserving DyNamic Transformer iv, 6, 7, 16, 17, 29, 31, 32, 34–37, 40–43, 58–63, 65

fps frames per second 49, 53

FVD fréchet video distance 40, 41, 43

GAN generative adversarial network 16

GCN graph convolutional network 9, 18, 21, 24, 28, 34

HAIKYU Hockey Action Identification and Keypose Understanding iv, 6, 7, 17–21, 28, 62, 63

HAR human action recognition iv, v, 1–3, 5, 16–18, 28, 29, 45, 48, 49, 53, 54, 56, 59, 62–64

HPE human pose estimation 8

LPIPS learned perceptual image patch similarity 40, 41, 43

PSNR peak signal-to-noise ratio 40, 41, 43

RGB red–green–blue color model 10, 13, 66

SSIM structural similarity index measure [40](#), [41](#), [43](#)

ST-GCN spatio-temporal graph convolutional network [9](#)

STGCN++ enhanced spatio-temporal GCN with multi-stream/adaptive features [9](#), [24](#), [34–36](#), [42](#), [59](#), [61](#)

ZoomGate scale-aware routing framework for mixed-zoom action recognition [v](#), [6](#), [7](#), [14](#), [17](#), [44](#), [45](#), [47–50](#), [53–55](#), [59](#), [61](#), [62](#), [64](#), [65](#)

List of Symbols

$H_{\max}$ Maximum bounding-box height over frames $1 \dots T$ [22](#)

T Number of frames in an action segment [22](#), [33](#)

$W_{\max}$ Maximum bounding-box width over frames $1 \dots T$ [22](#)

c_t^x, c_t^y Centre coordinates of the bounding box in frame t [22](#)

c_x, c_y General bounding-box centre coordinates [22](#)

w_t, h_t Width and height of the bounding box in frame t [22](#)

x_t, y_t Top-left pixel coordinates of the bounding box in frame t [22](#)

Top-1 Top-1 classification accuracy (%) [iv](#), [24](#), [42](#)

Chapter 1

Introduction

[Human action recognition \(HAR\)](#), the process of identifying and classifying human movements from visual data, has emerged as a crucial area of research in computer vision [2] due to its wide-ranging applicability across various domains, including sports analysis [49], autonomous driving [52], human-computer interaction [19], and video surveillance [23]. Yet accurate recognition remains challenging whenever shot scale changes, because the salient cues for an action shift with resolution: gross limb motion in wide views, torso twists in medium views, and facial or finger micro-expressions in close-ups. Moreover, reliable interpretation often hinges on [contextual information](#), which includes domain-specific signals such as sport equipment, scene layout, or specialized [keypoints](#) that disambiguate visually similar gestures.

The cinematic language further underscores the importance of scale agility. Directors routinely cut from a wide establishing shot that shows a full-body silhouette to a medium two-shot and then to a tight close-up, filling the frame with a single face or hand. Humans effortlessly track the continuity of motion across these radical changes in viewpoint and resolution, yet most learning-based recognition pipelines fail to do so. Each cut resamples action at a new pixel footprint, so the cues that matter shift from gross limb trajectories in the wide shot, to torso twists in the medium shot, to finger or facial micro-expressions in the close-up. A practical [HAR](#) system must therefore infer the current shot scale and dynamically prioritize the feature subset that remains reliable at that scale.

This thesis concentrates on two canonical yet contrasting settings: the [far-field scenarios](#), characterized by broadcast footage of sporting events that depict distant, rapid movements with lower-resolution visuals; and the [near-field scenarios](#), exemplified by close-up recordings of presenters, where subtle gestures and intricate details are visible. Address-

ing the unique challenges posed by these distinct scenarios requires the development of specialized methodologies tailored to their specific needs.

However, tackling either regime in isolation is insufficient once modern media (films, vlogs, instant replays) fluidly zooms or cuts between them. Consequently, we study skeleton-based HAR across the zoom continuum, with an emphasis on flexible, context-aware frameworks that stay robust as camera distance varies.

The remainder of this chapter (i) motivates the study in more detail, (ii) analyses the specific challenges of far-field scenarios, near-field scenarios, and zoom-mixed scenarios, (iii) states our research objectives, (iv) summarises the thesis contributions, and (v) provides a roadmap for the chapters that follow.

1.1 Motivation

HAR has witnessed significant advancements driven by increasing demand across diverse application areas. In sports analytics, precise action recognition can significantly enhance the analysis of player strategies, inform coaching decisions, and optimize athlete performance. For instance, identifying subtle hockey-specific movements can reveal strategic plays that are time-consuming to detect through manual observation.

At the opposite end of the camera spectrum, near-field scenarios, particularly the analysis and transformation of micro-gestures, hold potential for profound impacts in fields such as communication coaching, public speaking, and personal development. Effective recognition and editing of these subtle behaviours can help individuals present themselves with greater confidence and clarity. Figure 1.1 provides a visual illustration of the distinction between near-field scenarios and far-field scenarios, highlighting typical examples of each context and showcasing differences in visual conditions and keypoint visibility.

Generative video models further heighten the importance: realistic synthesis or editing depends on precise, frame-level action descriptors. When recognition fails, because a face suddenly becomes a tiny blob, or a full-body silhouette zooms into a close-up, the generative pipeline loses its guidance, producing implausible or identity-breaking artifacts. Robust, scale-aware HAR is therefore a prerequisite for trustworthy video generation and editing.

Despite their importance, existing research often focuses exclusively on either scenario, either in far-field scenarios, which are optimized for tiny, fast athletes under erratic camera motion, or in near-field scenarios, where high-resolution faces and hands dominate. Real media, however, rarely respects that boundary. Broadcast replays zoom from rink-wide

views to a shoulder-level crop of the goal scorer; vloggers cut between a head-and-shoulders camera and a wide room shot; autonomous-driving dashcams must recognize pedestrians who start as distant blobs and end as full-body close-ups next to the vehicle. A practical [HAR](#) engine must therefore continuously adapt its feature focus and representation as the apparent scale of the subject changes.

This thesis is motivated by the absence of a single, scale-aware framework that can simultaneously stay stable when athletes shrink to a few dozen pixels under drastic camera pans, remain sensitive when the very same subjects suddenly dominate the frame and reveal millimetre-scale gestures, and transition gracefully as the video stream zooms, cuts, or dollies between those extremes. To close that gap, we develop three complementary case studies. For the [far-field scenario](#) regime, we target ice-hockey broadcasts, where distant players, rapid motion, and stick-puck interactions stress existing methods. For the [near-field scenario](#) regime, we focus on presenter videos, detecting subtle confidence-related [micro-gestures](#). Finally, to cope with the [zoom-mixed scenarios](#) typical of cinematic editing, where directors jump from wide establishing shots to tight close-ups, we introduce a lightweight routing mechanism that fuses the specialist pipelines, yielding robust [HAR](#) across the full-scale continuum.

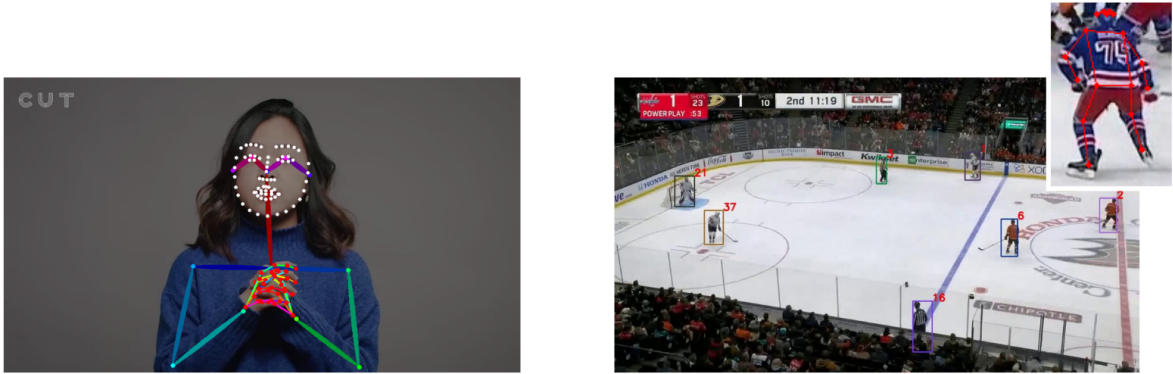


Figure 1.1: Illustration of near-field versus far-field action recognition scenarios, highlighting the distinct visual characteristics and challenges in each context. Near-field imagery emphasizes detailed, subtle gestures, which require fine-grained pose analysis, while far-field imagery typically includes distant, low-resolution subjects with frequent occlusions and rapid movement.

1.2 Challenges

Each camera scenario presents unique challenges that complicate traditional action recognition methods.

1.2.1 Far-field scenarios

Low resolution: Although broadcast footage may be high resolution overall, subjects typically appear very small in far-field views. This results in low-resolution representations of the players themselves, where detailed poses are difficult to estimate.

Camera-induced distortions: Rapid panning and zooming introduce spatial jitter, scale variations, and pose distortions, reducing action recognition accuracy.

Occlusions: Common in team sports such as ice hockey, occlusions often obscure critical body parts or equipment, making accurate skeletal keypoint extraction difficult.

Lack of sport-specific modelling: Traditional skeleton models omit essential elements such as ice-hockey sticks, which are crucial for distinguishing between similar actions.

1.2.2 Near-field scenarios

Fine-grained motion detection: Subtle [micro-gestures](#), such as finger fidgeting or crossed arms, require precise detection of minor movements, often missed by generic skeletal [keypoints](#).

Over-specification: High visibility can lead to excessive keypoint data, demanding efficient selection mechanisms to avoid unnecessary model complexity.

Partial occlusions: Hands or facial features may occasionally exit the frame, limiting data consistency and reducing recognition reliability.

Individual variability: [Micro-gestures](#) vary significantly across individuals and contexts, necessitating advanced methods for robust and meaningful classification.

1.2.3 Zoom-mixed scenarios

Shot-scale discontinuities: Abrupt cuts from a wide shot to a tight close-up (or vice-versa) break temporal smoothness; frame-to-frame [keypoints](#) trajectories become unreliable, confusing trackers and recurrent models.

Feature-relevance shift: Cues that are discriminative at one scale, such as whole-body silhouettes in wide views, become uninformative or invisible at another. In contrast, new cues, such as eyebrow raises in a close-up, suddenly appear. A single feature extractor must therefore re-weight its evidence on the fly.

These additional hurdles further reinforce the need for specialized [far-field scenarios](#) and [near-field scenarios](#) pipelines, as well as a fast routing mechanism that can mediate between them when the scale fluctuates within a single [zoom-mixed scenario](#).

1.3 Research Objectives

The primary objective of this thesis is to address the performance gap that arises when camera scale changes by devising scale-aware, skeleton-based pipelines for [far-field scenarios](#), [near-field scenarios](#), and [zoom-mixed scenarios](#). Concretely, we pursue five objectives:

1. **Robust distance-specific pipelines.** Design pose-centric recognition systems that remain reliable under (i) rapid, occlusion-prone sports footage and (ii) close-range videos dominated by closeup scale [micro-gestures](#).
2. **Domain-specific cues and normalization.** Introduce camera motion-compensated [bounding-box normalization](#) for [far-field scenario](#) tracks. Additionally, augment generic skeletons with task-critical key points, such as hockey-stick endpoints and whole-hand landmarks, for [near-field scenario](#) settings.
3. **Empirical benchmarking.** Build purpose-made datasets, train the proposed models, and provide a rigorous comparison against state-of-the-art baselines to quantify accuracy gains and failure cases.
4. **Scale-aware framework.** Develop a lightweight zoom-estimation and routing mechanism that dispatches each [tracklet](#) to the specialist recognizer best suited to its current shot scale.

Achieving these objectives pushes pose-based [HAR](#) toward a practical, context-aware solution that adapts seamlessly to camera distance.

1.4 Contributions

To address these objectives, this thesis delivers three projects, each targeting a different point on the zoom spectrum:

Research 1 (Far-field scenario): HAIKYU, addressing action recognition in far-field scenario hockey broadcasts through temporal bounding-box normalization and hockey-specific skeletons with stick endpoints.

Research 2 (Near-field scenario): CONFIDANT, focused on detecting and editing subtle near-field scenario micro-gestures to enhance perceived confidence, supported by a novel dataset and a diffusion-based generative editing pipeline.

Research 3 (Zoom-mixed scenario): ZoomGate, a scale-aware framework for cinematic video. It estimates the zoom level using the BB area and keypoints density, then routes tracklets to the appropriate specialist recognizer. Experiments show improved stability across heterogeneous zoom conditions, especially in film-style shot changes and continuous zooms.

1.5 Outline of the Thesis

This thesis explores skeleton-based action recognition across three contrasting visual domains: far-field scenarios sports broadcasts, near-field scenarios presenter videos, and zoom-mixed scenarios movie scenarios. It presents three core research projects, each forming the basis for a manuscript and contributing to a comprehensive understanding of pose-driven action recognition.

Chapter 1 introduces the research motivation and provides an overview of the thesis structure and objectives.

Chapter 2 surveys prior work in human action recognition with an emphasis on skeleton-based methods. It reviews challenges specific to far-field scenarios sports, near-field scenarios micro-gestures analysis, and zoom-mixed scenarios cinematic video, highlighting how scale variation complicates recognition. The chapter also discusses related advances in generative video editing, which provide context for the CONFIDANT framework, and establishes the need for scale-aware, context-sensitive recognition pipelines.

Chapter 3 outlines the methodological foundations used in far-field scenario sports scenarios, focusing on player tracking, pose extraction, and action classification in hockey broadcasts. It introduces HAIKYU: a pipeline designed to stabilize broadcast camera

effects and leverage hockey-specific keypoints for improved recognition performance. In addition, it evaluates the [HAIKYU](#) pipeline through both quantitative and qualitative analysis, including empirical results that compare keypoint configurations and normalization strategies, as well as visual examples and critical insights into limitations.

Chapter 4 describes the [CONFIDANT](#) framework for [near-field scenario micro-gestures](#) recognition and video-based behavioural editing. It details the creation of the CUT dataset, the pose-based generative editing pipeline, and the use of diffusion models for identity-preserving transformations. Additionally, it presents the experimental results of the [CONFIDANT](#) system, including accuracy in gesture classification, perceived confidence gains, and a qualitative evaluation of realism and continuity in edited videos.

Chapter 5 introduces [ZoomGate](#), a scale-aware framework designed for [zoom-mixed scenarios](#), where shot scales vary abruptly. [ZoomGate](#) employs heuristics based on [BB](#) area and [keypoints](#) density to estimate zoom level, then routes each person [tracklet](#) to the most appropriate specialist recognizer ([far-field scenario](#), [near-field scenario](#), or full-body). This approach ensures robust performance across film-style shot changes and continuous zooms. The chapter reports qualitative analyses that illustrate how explicit scale reasoning prevents failures in multi-scale video.

Chapter 6 concludes the thesis by summarizing the main contributions, identifying limitations across the three research projects, and outlining future directions. These include expanding datasets, developing unified architectures with scale and view invariance, and exploring multimodal fusion.

Chapter 2

Background Research

This chapter provides foundational knowledge and essential context required to understand the contributions and methodologies presented in this thesis. It covers key areas, including [human pose estimation \(HPE\)](#) and [skeleton-based action recognition](#), specific challenges associated with [far-field scenarios](#), [near-field scenarios](#), and [zoom-mixed scenarios](#), as well as context-aware recognition approaches. Finally, we outline advances in generative video editing that motivate our confidence-enhancing pipeline. Throughout this review, we highlight both state-of-the-art techniques and their limitations, laying the groundwork for the subsequent chapters.

2.1 Human Pose Estimation and Skeleton-Based Action Recognition

[Human pose estimation \(HPE\)](#) is the task of detecting key anatomical points, such as joints and limbs, on the human body in images or videos. [Skeleton-based action recognition](#) utilizes these structured pose data to represent human movements as sequences of [keypoints](#) configurations. By abstracting away appearance and background details, this representation simplifies the analysis of complex motion. Moreover, skeleton-based approaches exhibit robustness to variations in clothing and environmental conditions, which is especially valuable under occlusion or clutter [9].

Early methods typically employed generic two-dimensional skeleton models. A widely adopted standard is the [COCO keypoint](#) model, which defines 17 core body joint locations [31]. While this standard model provides broad applicability across various activi-

ties, it lacks the finer detail required in specialized domains such as gesture recognition or equipment sports.

To overcome this limitation, extended whole-body pose models have been introduced. Notably, the [COCO-WholeBody](#) dataset and model [27] annotates 133 [keypoints](#) per person, encompassing detailed landmarks for the face (68 points), hands (42 points), and feet (6 points), in addition to the core body joints. Figure 2.1 contrasts the standard and extended [COCO](#) configurations, illustrating that while the full-body model enables richer analyses of fine-grained gestures involving facial expressions or hand movements, it comes with increased complexity and computational cost.

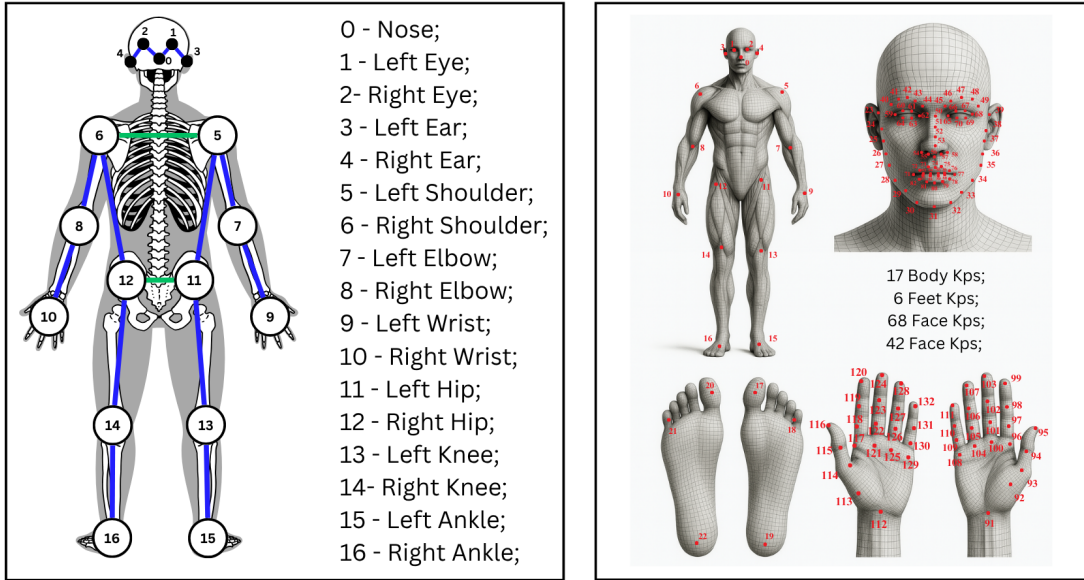


Figure 2.1: Comparison between the standard COCO skeleton model (left), comprising 17 body [keypoints](#), and the full-body COCO model (right), featuring additional [keypoints](#) for the face (68), hands (42), and feet (6). The extended model supports more comprehensive gesture analysis at the expense of increased computational complexity.

Recent advances in [skeleton-based action recognition](#) have shifted toward deep-learning frameworks, particularly [graph convolutional networks \(GCNs\)](#) methods that model the human body as a graph of joints. [spatio-temporal graph convolutional network \(ST-GCN\)](#) [55] and its variants treat each joint as a node in a graph and learn both spatial connections, such as bones or joint relations, and temporal dynamics of how poses evolve over time. Notable models in this line include [ST-GCN](#) [55], [2s-AGCN](#) [42], and [STGCN++](#) [14],

which have demonstrated high effectiveness in capturing complex movement patterns. Figure 2.2 illustrates a generic GCN-based action recognition workflow: a skeleton sequence is converted into a spatio-temporal graph, passed through stacked GCN layers, and finally mapped to an action-probability vector.

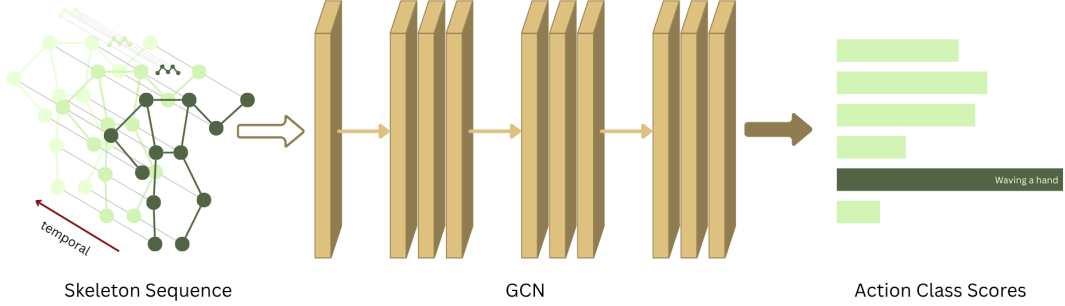


Figure 2.2: A video is first converted to a sequence of skeletons, which forms a spatio-temporal graph. Successive graph-convolution layers extract hierarchical motion features that a final softmax layer converts into class probabilities.

Although this thesis primarily addresses skeleton and [contextual information](#) due to their interpretability and efficiency, it acknowledges that incorporating other modalities, such as [red-green-blue color model \(RGB\)](#) video, [optical flow](#), audio, or inertial data, can further enhance recognition accuracy in specific contexts [25].

2.2 [Far-field scenario](#) Action Recognition

[Far-field scenario](#) action recognition involves understanding human actions when subjects appear at considerable distances from the camera, as commonly encountered in sports broadcasting and surveillance applications. This scenario presents unique challenges that are distinct from those in controlled, [near-field scenario](#) environments. In [far-field scenario](#) video footage, humans often appear small, typically only a few tens of pixels high. Such limited spatial resolution significantly reduces visual cues from limb articulation and motion. For instance, limb width in such scenarios may amount to merely a few pixels, thereby rendering pixel-based motion cues, such as [optical flow](#), unreliable [9].

Figure 2.3 (typical of broadcast footage, where athletes appear as tiny targets) compares the relative bounding-box sizes of players in ice hockey and volleyball games. Even though both frames have the same overall pixel dimensions (1280×720), a hockey skater’s

bounding box typically spans 6,780–17,010 pixels, whereas a volleyball player’s covers about 16,800–19,323 pixels. Both frames have the same overall resolution ($1280 \times 720 = 921,600$ pixels), but far-field subjects occupy a much smaller pixel footprint, making limb details blur into just a handful of pixels and motion cues harder to capture.

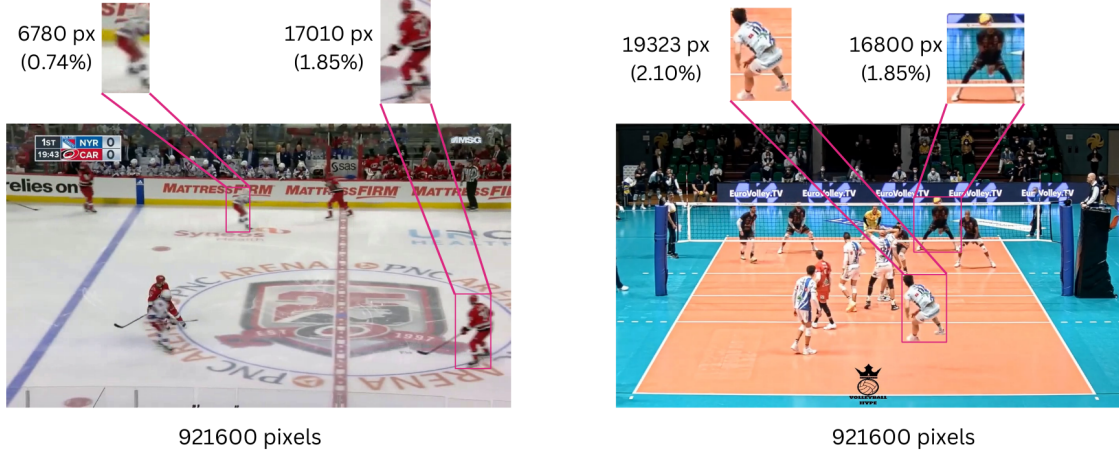


Figure 2.3: Broadcast camera placement often renders athletes as tiny targets. In hockey (left), each player’s bounding box spans only 6,780–17,010 pixels, compared to 16,800–19,323 pixels in volleyball (right), even though both frames share the same 1280×720 resolution. Such small footprints make [far-field scenario](#) action recognition markedly harder.

Further complicating [far-field scenario](#) recognition is the frequent camera motion (panning and zooming), which induces motion blur and background variations. Concurrently, crowded scenes and player occlusions introduce ambiguities. Reliable recognition often depends on tracking interactions with sports equipment, such as distinguishing between a casual arm movement and a tennis swing based on the presence and use of a racket. Thus, temporal stabilization, player tracking, contextual modelling, and object detection are often used to address these challenges.

Despite this progress, [far-field scenario](#) action recognition remains a complex problem. Datasets such as Sports-1M [28] and SoccerNet [20] have contributed valuable benchmarks and driven algorithmic improvements. However, as Host et al. [25] highlight in their survey, even state-of-the-art deep learning architectures struggle with the combined complexity of small subject scale, fast motion, crowded environments, and unstable camera views. Therefore, achieving robust recognition under these conditions remains an open and active

area of research.

2.3 Near-field scenario Action Recognition

Near-field scenario action recognition focuses on scenarios where the subject is close to the camera, often resulting in high-resolution views of upper-body movements, facial expressions, and hand gestures. Examples include interpreting a person’s gestures and body language during a presentation or conversation, where subtle movements can convey significant meaning. In such settings, the goal is frequently fine-grained gesture detection rather than large-scale actions. Of particular interest to this thesis is the interpretation of emotional or psychological states (such as confidence or anxiety) through **micro-gestures** and body language. These cues are typically much more subtle than overt actions, for instance, brief self-comforting gestures such as fidgeting with glasses or adjusting one’s hair. Recognizing these reliably demands high accuracy and often specialized models, since **micro-gestures** can be easily overshadowed by larger movements or may not repeat frequently. Figure 2.4 illustrates three representative **near-field scenario** examples. Each frame contains an overlaid whole-body pose estimation, highlighting how dense facial, hand, and torso **keypoints** capture the subtle **micro-gestures** that are the focus of this work.

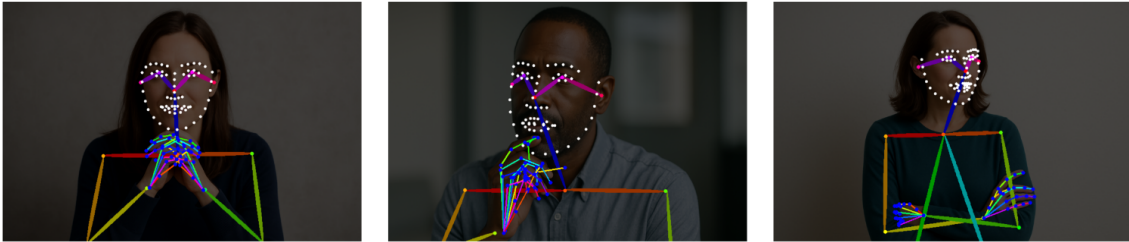


Figure 2.4: Three **near-field scenario** samples with upper-body pose estimations. The dense facial and hand landmarks emphasize why **near-field scenario** recognition benefits from fine-grained, whole-body **keypoints**.

Body language research indicates that confident body language, such as open postures and deliberate gestures, enhances credibility, whereas unconfident cues such as slouching or fidgeting diminish perceived authority [29]. Thus, systematically recognizing and enhancing these subtle cues has considerable potential to improve communication effectiveness, particularly in video presentations.

Currently, publicly available datasets specifically targeting confidence-related body language are sparse. Existing [micro-gestures](#) datasets, such as the Spontaneous Micro-gesture (SMG) dataset [10] and the In-the-wild Micro-Gesture Understanding and Emotion Analysis (iMiGUE) dataset [32], focus primarily on emotional states and spontaneous gestures. The SMG dataset, for instance, annotates gestures under stressed and non-stressed conditions, whereas iMiGUE captures spontaneous gestures from athletes in high-stakes environments.

Despite their value, these datasets highlight the gap in resources explicitly geared towards confidence-related gestures. This limitation highlights the need for developing a comprehensive dataset specifically annotated for both confident and unconfident behaviours, thereby facilitating improved gesture analysis in communication scenarios.

2.4 [Zoom-mixed scenario](#) Action Recognition

[Zoom-mixed scenario](#) action recognition requires a model to stay reliable while the on-screen size of the actor expands or contracts from long shots to extreme close-ups. Each scale privileges a different cue hierarchy: whole-body kinematics dominate when the person is small, whereas subtle hand motions and facial [micro-gestures](#) become decisive once the camera zooms in. Figure 2.5 visualizes the five canonical shot scales provided in MovieNet, underscoring how radically the available cues change across them.

Early attempts at scale-robust recognition include COMpuTER [46] and Structured Keypoint Pooling [22]. COMpuTER blends skeletal and [RGB](#) streams via human-centric queries and a cross-modal consistency loss, prioritizing [keypoints](#) when resolution is high and reverting to appearance cues when the actor becomes small. Structured Keypoint Pooling merges human joints and object trajectories into a single three-dimensional [point cloud](#), pooling features hierarchically so that detection noise at either scale has less impact. Both systems improve tolerance to camera distance; yet every frame still traverses the same extensive backbone, and the choice of scale remains implicit rather than explicit.

A complementary line of work is Shahabinejad et al. [41]. Their method inserts a differentiable crop module ahead of a standard 3-D CNN and trains it end-to-end, using compressed-domain motion residuals to learn where and how far to zoom. The policy adds negligible computational cost and reliably boosts accuracy, proving the value of learning a frame-wise zoom operation. Nevertheless, it operates only in the [RGB](#) domain and does not adapt skeletal resolution or cue weighting once the crop is selected.

A natural testbed for these ideas is MovieNet [26], a corpus of 1,100 full-length films

whose 92k cinematic-style tags annotate every shot with its camera scale (extreme close-up, close-up, medium, full, long) and motion (zoom-in, zoom-out, static). Combined with 65k action labels, MovieNet enables evaluation splits in which scale flips repeatedly within a scene, precisely the scenario that stresses [zoom-mixed scenario](#) recognition. Generic datasets such as Kinetics or Sports-1M contain scale variation only implicitly and lack shot metadata, making controlled testing significantly more challenging.

Taken together, current work leaves three issues open: (i) rapid, reliable detection of scale shifts; (ii) dynamic re-weighting of skeletal versus appearance cues; and (iii) compute allocation that matches the present zoom. The [ZoomGate](#) framework, introduced in Chapter 5, addresses all three by computing a lightweight [scale gate](#), switching to a scale-specific skeleton, and routing every [tracklet](#) to the [far-field scenario](#) or [near-field scenario](#) expert best suited for that moment.

2.5 Contextual and Object-Aware Action Recognition

While [skeleton-based action recognition](#) models primarily analyze human joint movements, incorporating [contextual information](#) and interactions with objects significantly enhances recognition accuracy. Additional cues beyond pure skeletal dynamics help disambiguate visually similar actions that differ in meaning. Recent studies have demonstrated that explicitly modelling temporal context and object interactions can substantially improve classification performance, particularly in challenging scenarios.

Cioppa et al. introduced a Context-Aware Loss Function (CALF) [12] designed for action spotting in soccer footage. Instead of predicting actions based solely on isolated moments, CALF penalizes incorrect predictions over an extended temporal window surrounding each event. By accounting for preceding and subsequent events, this method accurately distinguishes similar actions, such as differentiating a scored goal from a missed attempt based on subsequent frames showing either celebration or continued gameplay. Experiments on the SoccerNet dataset confirmed substantial gains in recognition accuracy, illustrating the importance of temporal context.

Additionally, object interactions and environmental contexts provide essential cues for distinguishing actions with similar skeletal movements. Wen et al. [51] proposed integrating object nodes directly into skeleton graphs, allowing for the explicit modelling of interactions between human joints and relevant objects, such as a hand joint connected to a cellphone, indicating texting. This method significantly improved recognition performance in object-involved activities and clarified ambiguous skeletal gestures. Thus, contextual

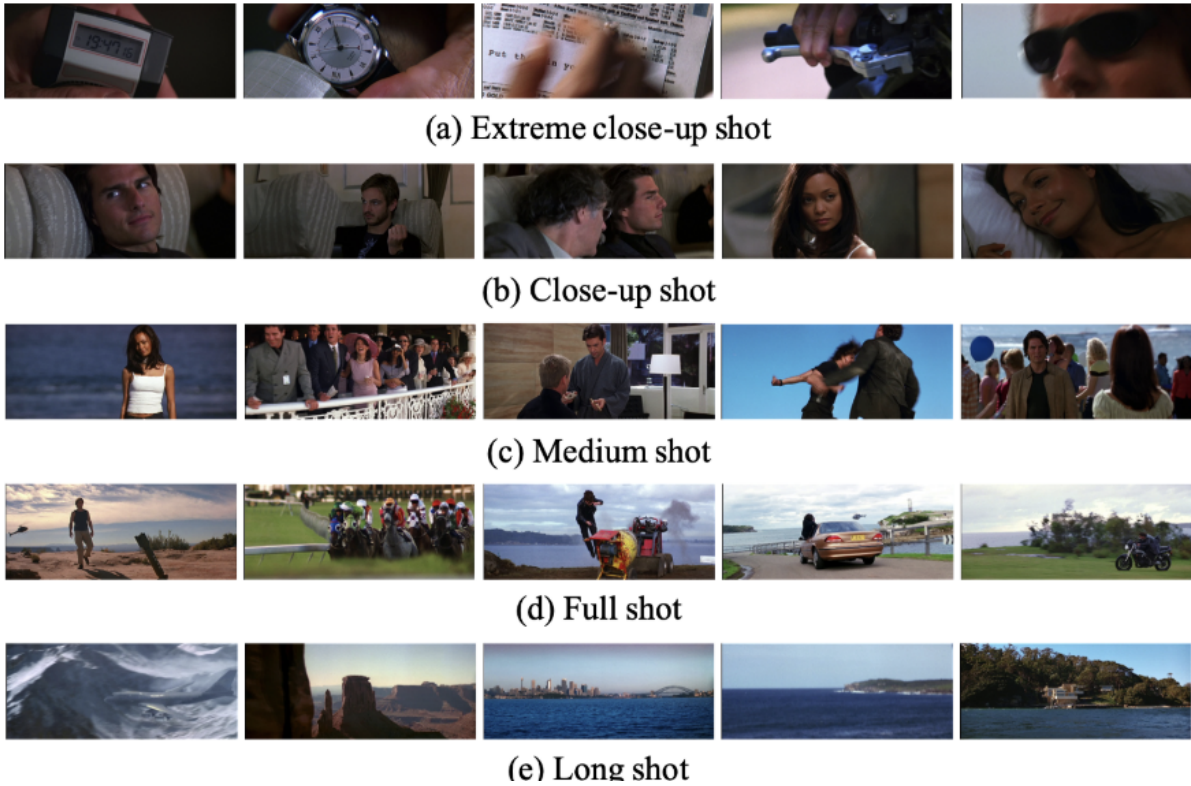


Figure 2.5: Examples of the five canonical shot scales annotated in MovieNet: (a) extreme close-up, (b) close-up, (c) medium, (d) full, and (e) long shot. The drastic change in actor size across rows highlights the need for explicit scale reasoning in *zoom-mixed scenario* action recognition. *Source: MovieNet [26].*

and object-aware approaches enable richer understanding of actions, highlighting the limitations of relying exclusively on skeletal data.

2.6 Generative Video Editing

Recent developments in video generative models have shown impressive abilities in producing high-quality video content [57, 6]. Techniques such as [generative adversarial networks \(GANs\)](#) [37] have enabled the generation of realistic video sequences and style transformations. DeepFake [35] technologies represent a significant advancement in this field, enabling realistic face swapping and expression manipulation; however, they often compromise individual identity, making them unsuitable for applications where preserving the speaker’s authenticity and identity is crucial.

Another approach is MimicMotion [59], which introduces motion pattern transfer between subjects, enabling the replication of specific movements and gestures. While this approach illustrates the potential for modifying body movements in videos, it may not account for the preservation of individual identity or the subtle [micro-gestures](#) that convey meaning. Preserving subtle [micro-gestures](#) is essential because these small, often involuntary movements convey significant information about a person’s emotional state and confidence levels. Altering them without careful consideration can lead to unnatural results and diminish the authenticity of the presentation.

Concurrently, transformer-based [HAR](#) backbones such as TimeSformer [7] and VideoMAE [44] set strong RGB baselines via long-range spatiotemporal attention and masked autoencoding pretraining. Our scale-aware, pose-driven pipeline is complementary: instead of relying on a single heavy appearance backbone, we explicitly detect shot scale and reweight skeleton vs. appearance cues, so these transformers can be plugged in as near-/far-field experts within ZoomGate while retaining our routing and compute allocation.

In the context of this thesis, generative video editing is considered a downstream task; its effectiveness depends directly on the accuracy and granularity of the upstream [HAR](#) pipeline. The [CONFIDANT](#) framework (see Chapter 4) leverages fine-grained, pose-based action recognition to guide generative editing, ensuring that confidence-enhancing transformations are grounded in reliable gesture detection. By framing editing as a downstream application, we highlight how improvements in [HAR](#) at varying camera scales directly enable higher-quality and more identity-preserving video synthesis.

2.7 Conclusion

The literature review has exposed three recurring pain points. First, [far-field scenario](#), [near-field scenario](#), and [zoom-mixed scenario](#) footage each demand a different skeleton granularity and cue hierarchy, yet most existing models treat them uniformly. Second, [contextual information](#) and object interactions can disambiguate actions but are rarely fused with pose in a scale-aware manner. Third, generative editors can alter body language, but few safeguard subtle [micro-gestures](#) or identity.

The following three chapters tackle these gaps in turn. [HAIKYU](#) augments hockey poses with [stick keypoints](#) to stabilize [far-field scenario](#) recognition; [CONFIDANT](#) focuses on [near-field scenario micro-gestures](#) and demonstrates identity-preserving confidence editing; and [ZoomGate](#) unifies the spectrum by detecting scale on the fly, switching skeleton resolution, and routing each [tracklet](#) to the specialist best suited for that moment. Taken together, these contributions advance [HAR](#) toward a robust, scale-aware, and contextually grounded pipeline.

Chapter 3

Methodology (Far-Field Scenario: HAIKYU)

Player action recognition is a fundamental task in sports analytics, enabling the automatic identification and classification of particular player movements during the game, such as skating, striking a puck, or maintaining position. Understanding these actions allows for more profound insights into player strategies and game analysis. Although sports such as basketball [11, 30] and soccer [60, 39] have seen extensive development in player action recognition methodologies due to their popularity, ice hockey remains comparatively underexplored. Traditional approaches, such as the Action Recognition Hourglass Network (ARHN) [18], primarily rely on static pose input, thereby ignoring temporal and [contextual information](#). By [contextual information](#) of broadcast ice hockey videos, we refer to structured domain-specific cues, such as camera-driven scale changes and [stick keypoints](#), that extend beyond joint coordinates. The two contextual priors that we leverage in this work are visualized in Figure 3.1.

This chapter presents the detailed methodology behind [HAIKYU](#), a framework designed specifically for [far-field scenario HAR](#) within ice hockey broadcast footage. Ice hockey presents unique challenges, including rapid player movements, frequent occlusions due to protective gear and stick/puck interactions, and significant variability in player scale resulting from dynamic camera movements. Addressing these challenges is crucial for accurate event detection and player performance analytics. In this chapter, we describe our specialized Ice Hockey dataset, elaborate on the technical pipeline (encompassing player tracking, [bounding-box normalization](#), hockey-specific pose extraction, and [GCN](#)-based classification), provide comprehensive quantitative and qualitative analyses, and offer crit-

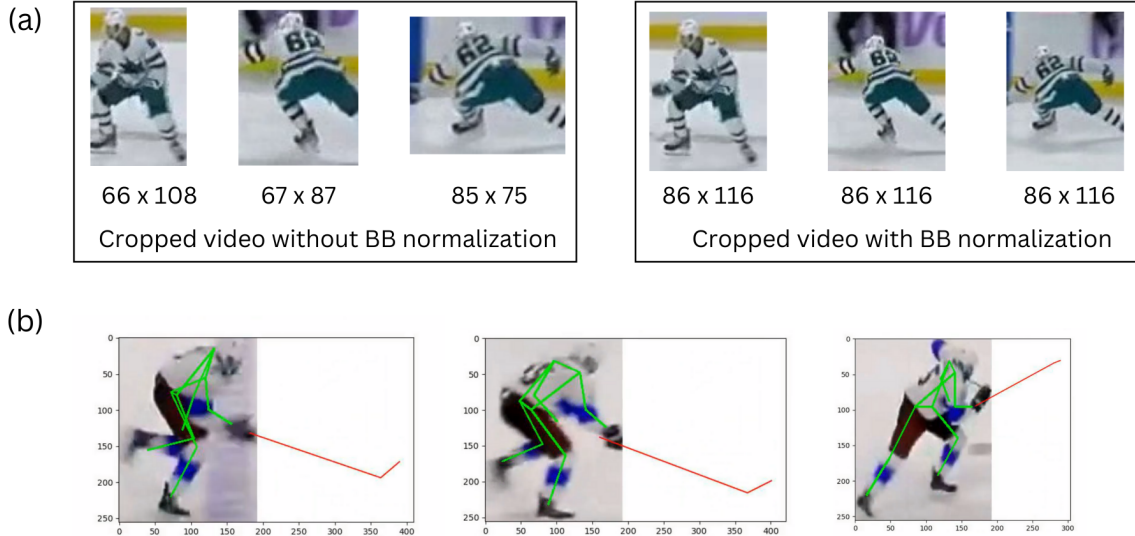


Figure 3.1: Illustration of the two contextual priors exploited by **HAIKYU**. (a) Temporal **bounding-box normalization** suppresses scale jitter by resizing every frame in the player **tracklet** to the maximum **BB** dimensions. (b) A 15-**keypoints** ice hockey skeleton (green body joints, red stick joints) with **stick keypoints**, which are vital for player action recognition and not included in the generic 17-**COCO** skeleton.

ical insights into the impact of hockey-specific **keypoints** configurations and contextual priors introduced by our approach.

3.1 Ice Hockey Dataset

The Ice Hockey dataset was developed to facilitate action recognition in challenging **far-field scenario** broadcast video scenarios, specifically ice hockey games. It comprises annotated video clips collected from National Hockey League (NHL) broadcasts recorded at 30 frames per second (fps). The dataset consists of a variety of clips sourced from multiple games, encompassing 29 teams and featuring diverse visual conditions, including different arenas, lighting, uniforms, and camera angles.

Each annotated action instance in the dataset spans precisely 2 seconds (60 frames), centred around critical action frames, and the annotation was performed manually. The

dataset features six specific ice hockey actions. Table 3.1 summarizes the class definitions and their distribution.

Table 3.1: Action classes with descriptions and frequency.

Action Class	Description	Occurrences
Rapid Deceleration	Player abruptly reduces skating speed to change direction or avoid contact	52
Backwards Skating	Player skates in reverse, typically to maintain defensive positioning	116
Maintaining Position	Player remains largely stationary while reading the play or screening the goalie	164
Strike Puck with Stick	Player’s stick makes direct contact with the puck for a pass, shot, or block	337
Forwards Skating	Player accelerates toward the offensive zone while controlling balance and speed	421
Stickwork	Player performs intricate stickhandling to retain or regain puck control	457
Total		1547

The dataset comprises 1,547 annotated action instances across six classes. The distribution is imbalanced, with common actions such as Stickwork (457) and Forwards Skating (421) far more frequent than Rapid Deceleration (52). Because of this imbalance, reporting only overall accuracy would be misleading; therefore, in later experiments, we additionally present confusion matrices to show per-class performance. To prepare the dataset, we extracted player [tracklets](#) using the VIP-HTD [38] tracking method. Additionally, [bounding-box normalization](#) was applied across temporal segments to stabilize player scales, significantly reducing spatial jitter. A formal dataset note appears in Appendix A.

3.2 Methodology

3.2.1 Pipeline Overview

The [HAIKYU](#) pipeline for ice-hockey action recognition comprises four stages that together inject two complementary [contextual information](#) priors: a temporal prior that stabilizes

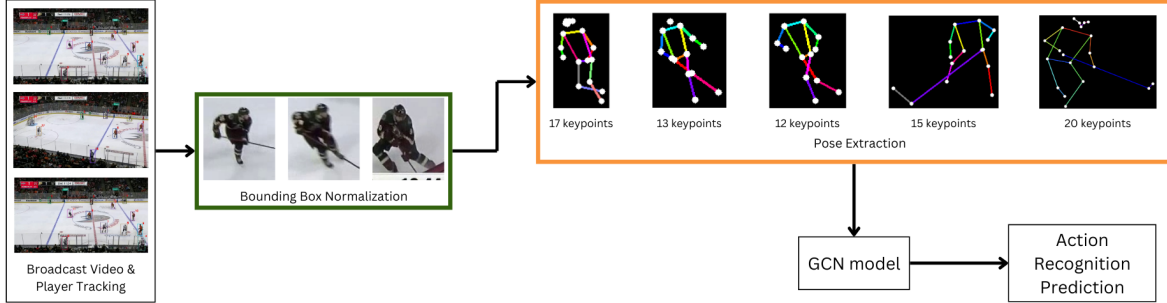


Figure 3.2: Overview of the [HAIKYU](#) pipeline. Contextual Prior #1 (green) stabilizes [BB](#) scale; Prior #2 (orange) augments the skeleton with [stick keypoints](#). The enriched graph is fed to a [GCN](#)-based model for action recognition.

player scale and position, and an object prior that models the hockey stick explicitly. Figure 3.2 gives a high-level view of the proposed pipeline.

Each annotated action segment is first preprocessed by automatically extracting a player [tracklet](#) with the VIP-HTD tracker [38]. [BBs](#) are then temporally normalized to enforce scale consistency (Section 3.2.3, Contextual Prior #1). Next, an ice-hockey-specific pose estimator [4, 5] converts every normalized crop to a skeleton, yielding 2-D [keypoints](#) sequence with six alternative configurations (Section 3.2.4, Contextual Prior #2). Finally, a [GCN](#)-based model classifies the action (Section 3.2.5).

3.2.2 Preprocessing

To automate [BB](#) extraction and ensure temporal consistency across frames, we employ a tracking approach [38], tailored explicitly for player detection in ice hockey broadcast footage. After extracting [BBs](#), frame-wise cropping and resizing are performed to a normalized spatial representation. These cropped [BBs](#) serve as inputs for pose extraction, which utilizes ice-hockey-optimized models presented in [4, 5].

3.2.3 Bounding Box Normalization

A significant challenge in using broadcast video for pose estimation is the variability in player size and position across frames. A naive approach might resize each frame independently, stretching the player’s skeleton and distorting body shape, which degrades

downstream performance. To mitigate this, we propose a robust **bounding-box normalization** strategy. Specifically, for each annotated action we compute the maximum width $W_{\max}$ and height $H_{\max}$ of the BB across the segment, then re-extract crops so every frame uses these maxima while keeping the player centred.

Notation. A player **tracklet** contains T frames. In frame t , the bounding box has top-left pixel coordinates (x_t, y_t) and dimensions (w_t, h_t) . We denote the box centre by (c_t^x, c_t^y) .

We first compute the segment-wise maxima:

$$W_{\max} = \max_{1 \leq t \leq T} w_t, \quad H_{\max} = \max_{1 \leq t \leq T} h_t. \quad (3.1)$$

For each frame t , we centre a box of size $(W_{\max}, H_{\max})$ at the original centre:

$$\begin{aligned} c_t^x &= x_t + \frac{w_t}{2}, & c_t^y &= y_t + \frac{h_t}{2}, \\ x'_t &= c_t^x - \frac{W_{\max}}{2}, & y'_t &= c_t^y - \frac{H_{\max}}{2}. \end{aligned} \quad (3.2)$$

Here, (c_t^x, c_t^y) are the bounding-box centre coordinates and (x'_t, y'_t) are the top-left coordinates of the normalized box. This ensures a consistent BB size and centre throughout the segment, suppressing camera-induced scale jitter.

To illustrate the effect, Figure 3.3 compares **keypoint** trajectories without and with normalization. Without normalization, trajectories are scattered due to scale changes; after normalization, they are markedly more stable and coherent.

3.2.4 Pose Keypoint Extraction

We apply ice-hockey-optimized 2D pose estimators [5, 4] to generate skeletal **keypoints** from players in normalized BBs belonging to a sequence. We experimented with various **keypoint** configurations (Figure 3.4):

1. **17 Keypoints (COCO [31]):** Standard full-body representation.
2. **13 Keypoints:** Merged head **keypoints**.
3. **12 Keypoints:** Head **keypoints** omitted to focus solely on torso and limbs.
4. **15 Keypoints:** 12 body **keypoints** plus 3 **stick keypoints**.

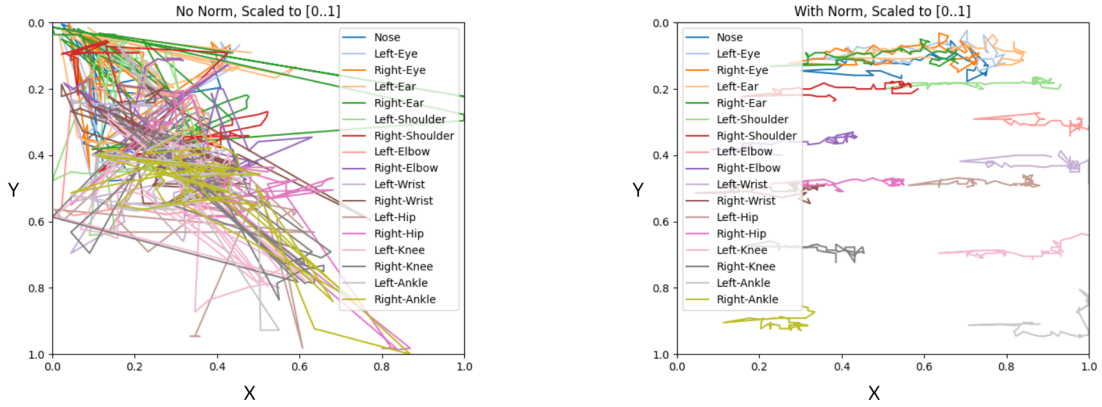


Figure 3.3: Impact of [bounding-box normalization](#) on [keypoint](#) trajectories for one data sample. Left: Without normalization, trajectories are scattered and inconsistent due to variability in player scale. Right: After applying [bounding-box normalization](#), trajectories appear significantly more stable and coherent, clearly reducing positional variance and making movement patterns easier to interpret.

5. **20 Keypoints:** Comprehensive setup with 5 head, 12 body, and 3 [stick keypoints](#).
6. **3 Stick Keypoints Only:** Focused solely on stick interaction.

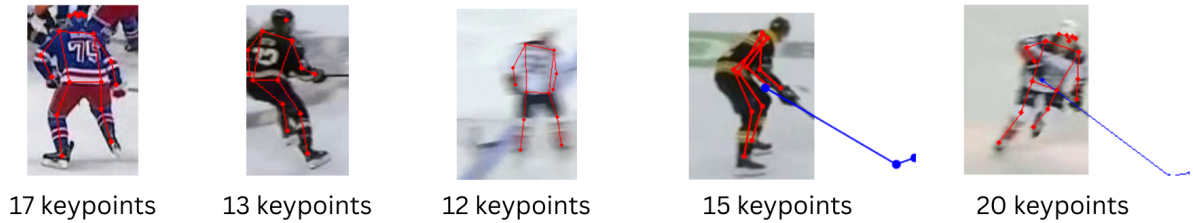


Figure 3.4: Five skeletal configurations used in this study, from left to right: 17-[COCO keypoints](#), 13-keypoint merged head, 12-keypoint head-omitted, 15-keypoint with [stick keypoints](#), 20-keypoint comprehensive configuration.

3.2.5 GCN for Action Recognition

We employ the [two-stream adaptive graph convolutional network \(2s-AGCN\)](#) [42] and [enhanced spatio-temporal GCN with multi-stream/adaptive features \(STGCN++\)](#) [17] architectures to classify skeleton-based action sequences. Both are state-of-the-art graph convolutional models designed for action recognition, where joints are treated as nodes and their spatial-temporal relationships are modeled as edges. We selected these architectures for two reasons. First, they have consistently demonstrated strong benchmark performance across large-scale skeleton datasets, making them reliable baselines for pose-driven recognition. Second, their ability to adaptively weight joint connections and model temporal dependencies aligns well with the challenges of our hockey dataset, where rapid motion, occlusion, and small player scale can obscure key joint trajectories. Training and evaluation leverage the mmAction2 toolbox [13]. Our approach exclusively utilizes the joint modality ([keypoints](#) 2D coordinates) to capture spatial-temporal dynamics. This choice simplifies model inputs while effectively capturing both static postures and dynamic movements necessary for accurate ice hockey action recognition.

3.3 Experiments

3.3.1 Experimental Setup

To ensure a robust model evaluation, we partitioned the dataset described in Section 3.1 into training and validation sets using a 70%–30% stratified split, maintaining proportional representation of each action class. We evaluate action recognition performance using standard metrics, reporting [Top-1](#) accuracy unless otherwise stated.

We adopt the mmAction2 toolbox [13] for training and evaluation. The training process leverages the [GCN](#)-based architectures (described in Section 3.2.5) with cross-entropy loss. Experiments were conducted on an NVIDIA RTX 3090 Ti GPU. The training setup for our GCN experiments is as follows: a learning rate of 0.1, weight decay of 5×10^{-4} , batch size of 16, and a total of 20 epochs, using stochastic gradient descent (SGD) as the optimizer.

3.3.2 Experimental Results

We evaluated our ice hockey action recognition framework using exclusively the joint modality, focusing on different skeletal configurations and their impact on performance. Results across these configurations are presented in Table 3.2.

Table 3.2: Comparison of action recognition accuracy using joint modality across different skeletal configurations and models.

Model	Keypoint Configuration	Accuracy (%)
STGCN++	17 Keypoints (No bounding-box normalization)	31
	12 Keypoints	40
	13 Keypoints	44
	17 Keypoints	57
	15 Keypoints (12 body + 3 stick)	<u>64</u>
	20 Keypoints (5 head + 12 body + 3 stick)	65
	3 Stick Keypoints Only	52
2s-AGCN	17 Keypoints	51
	15 Keypoints (12 body + 3 stick)	56

The 20-keypoint skeleton with STGCN++ attained the highest accuracy (65%). Using the same model, adding [bounding-box normalization](#) to the standard 17-COCO skeleton raised accuracy from 31% to 57%. Adding three [stick keypoints](#) increased accuracy by a further 7 percentage points to 64%, and supplementing with five head landmarks reached 65%. Conversely, dropping head landmarks (13 or 12 keypoints) lowered accuracy to 44% and 40%, respectively. **These stepwise gains and losses confirm that both detailed head cues and explicit stick endpoints are essential for recognizing hockey-specific actions.** Additionally, the three-stick-keypoint-only configuration (52%) highlights the significant role of stick interactions. Overall, STGCN++ consistently outperformed 2s-AGCN, validating its selection as the core recognizer.

The matrices demonstrate that STGCN++ (20 keypoints) achieves the highest accuracy, distinguishing subtle actions such as stickwork versus striking the puck. In contrast, models without explicit [stick keypoints](#) or [bounding-box normalization](#) frequently misclassify these actions, underscoring the significance of detailed skeletal representations and spatial normalization.

3.3.3 Qualitative Results

Figure 3.6 illustrates qualitative predictions from our models on six representative clips. Each image captures the midpoint frame of a two-second sequence, along with ground-truth labels and predictions from five setups: (1) without [bounding-box normalization](#), (2) 2s-AGCN with 17 keypoints, (3) STGCN++ with 15 keypoints, (4) STGCN++ with 17 keypoints, and (5) STGCN++ with our optimized 20-keypoint configuration.

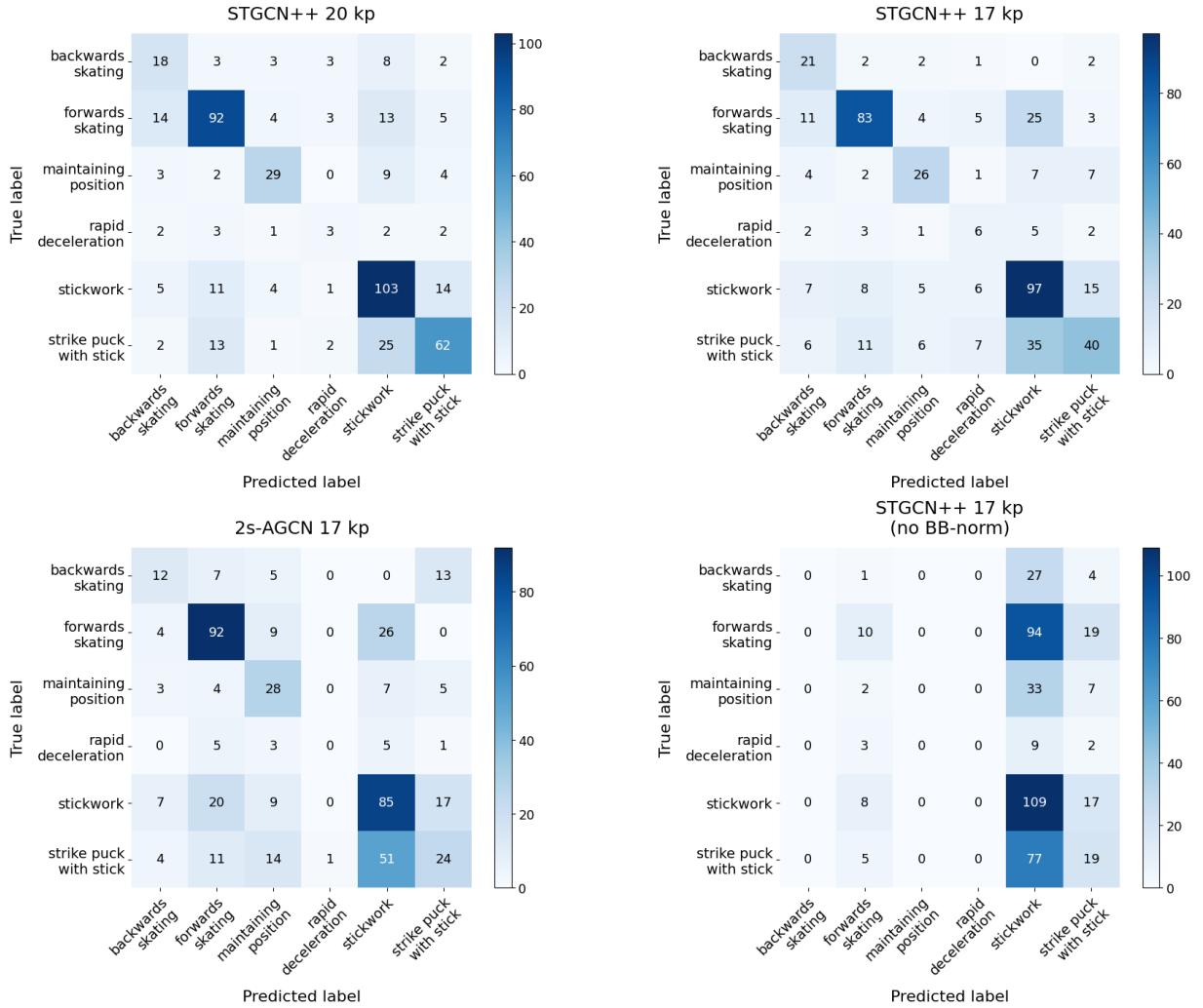


Figure 3.5: Confusion matrices comparing action recognition performance across different models and [keypoint](#) configurations. From left to right: STGCN++ with 20 keypoints (best performing), STGCN++ with 17 keypoints, 2s-AGCN with 17 keypoints, and STGCN++ with 17 keypoints without [bounding-box normalization](#). Higher diagonal values indicate better accuracy, while off-diagonal entries represent common misclassifications.

Without [bounding-box normalization](#), subtle actions are often misclassified, with stick interactions mistaken for generic stickwork due to spatial inconsistencies. The 2s-AGCN baseline with 17 keypoints struggles to differentiate stick actions from skating movements.

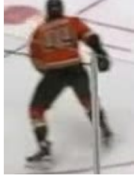
						
Ground Truth	Strike puck with stick	Forwards skating	Backwards skating	Maintaining position	Rapid deceleration	Stickwork
Our Prediction (w/o bb norm)	Stickwork	Stickwork	Stickwork	Strike puck with stick	Stickwork	Stickwork
Our Prediction (2s-AGCN 17 kps)	Stickwork	Forwards skating	Strike puck with stick	Maintaining position	Stickwork	Strike puck with stick
Our Prediction (STGCN++ 15 kps)	Strike puck with stick	Forwards skating	Forwards skating	Maintaining position	Stickwork	Stickwork
Our Prediction (STGCN++ 17 kps)	Strike puck with stick	Forwards skating	Stickwork	Strike puck with stick	Stickwork	Stickwork
Our Prediction (STGCN++ 20 kps)	Strike puck with stick	Forwards skating	Backwards skating	Maintaining position	Stickwork	Stickwork

Figure 3.6: Qualitative comparison of ice hockey action recognition predictions across different setups on six clips. Ground-truth labels are shown in green (correct predictions), while incorrect predictions are shown in red. The optimized STGCN++ model with 20 keypoints demonstrates improved accuracy, effectively differentiating subtle actions compared to models without [bounding-box normalization](#) or fewer keypoints.

STGCN++ with 15 or 17 keypoints improves accuracy, while the optimized 20-keypoint configuration excels, reliably distinguishing forward versus backward skating and “strike puck with stick” versus general stickwork. These findings highlight the importance of incorporating detailed head and stick keypoints together with effective [bounding-box normalization](#).

While our optimized STGCN++ configuration achieved notable gains over the baselines, these accuracies still fall short of the level required for operational deployment in professional sports analytics. This gap stems from several factors: the dataset is limited in size and imbalanced across classes, restricting generalization; the models rely solely on skeletons, without multi-view or optical flow cues to disambiguate similar motions; and the lack of puck tracking hampers differentiation between actions such as stickwork and striking. Future improvements could come from larger, balanced datasets, integrating contextual cues and reliable puck trajectories, and advancing toward 3D pose estimation and

multi-person context modelling to reduce misclassifications.

3.4 Conclusion

This chapter presented [HAIKYU](#), a domain-specific pipeline for [far-field scenario HAR](#) in broadcast ice hockey. The method integrates two complementary [contextual information](#) priors: (i) a temporal prior, implemented as [bounding-box normalization](#) to suppress camera-induced scale jitter, and (ii) an object prior, implemented through hockey-specific [stick keypoints](#) and optional head landmarks.

Comprehensive experiments on the Ice Hockey dataset demonstrated that these priors substantially improve classification accuracy under challenging broadcast conditions. [bounding-box normalization](#) alone raised accuracy from 31% to 57% with a 17-[COCO](#) skeleton, while adding three stick keypoints increased performance to 64%. The full 20-keypoint configuration, combining stick endpoints with fine-grained head landmarks, achieved the highest accuracy of 65%. Stepwise analysis confirmed that gains derive not from simply adding more keypoints, but from selecting contextually relevant ones aligned with the sport’s semantics.

Qualitative results further highlighted the benefits of these design choices, improving disambiguation between visually similar actions such as stickwork and striking the puck, as well as forward versus backward skating. These findings underscore the importance of both spatial normalization and domain-specific skeletal modelling for reliable [far-field scenario](#) recognition.

While [HAIKYU](#) was validated on pre-segmented clips, extending it to continuous video for automated action spotting remains an open challenge. Future work could also explore multi-view or depth-based pose estimation for enhanced 3D accuracy, and richer cues such as puck tracking or modelling inter-player interactions to capture tactical plays.

By combining robust temporal stabilization, sport-specific [keypoints](#), and spatio-temporal [GCN](#) modelling, [HAIKYU](#) advances the state of [far-field scenario](#) recognition in equipment-centric sports. The insights gained here provide a transferable blueprint for adapting pose-centric [HAR](#) pipelines to other domains with similar challenges.

Chapter 4

Methodology (Near-Field Scenario: CONFIDANT)

Building on the thesis-wide motivation of [scale-aware routing](#) in [HAR](#), this chapter focuses on the [near-field scenario](#) setting, where subjects appear close to the camera and subtle upper-body movements become the dominant source of communicative cues. In this context, we introduce [CONFIDANT](#), a framework for the fine-grained recognition of confidence-related [micro-gestures](#) such as folded arms, finger fidgets, and clasped hands in presenter-style videos. Unlike large-scale bodily actions in sports or cinematic footage, these [micro-gestures](#) are fleeting, low-amplitude, and often overlooked, yet they are crucial for interpreting perceived confidence. Accurate detection of such cues not only supports scientific understanding of non-verbal behaviour but also enables downstream applications; one such example, identity-preserving video editing, is demonstrated in [Section 4.3](#), though the primary contribution of this chapter remains the recognizer itself.

4.1 CONFIDANT Dataset

4.1.1 Existing Micro-Gesture Datasets

A significant limitation in advancing [near-field scenario](#) gesture recognition, particularly in terms of confidence or affective states, is the scarcity of labelled datasets. Most publicly available action/gesture datasets focus on macro-actions or standard emotion recognition, rather than the nuanced behaviours linked to confidence. Two notable datasets in the

realm of [micro-gestures](#) are SMG [10] and iMiGUE [32], although their emphasis is slightly different.

The Spontaneous Micro-Gesture (SMG) dataset was introduced to study upper-limb [micro-gestures](#) under different emotional stress states. It contains 3,692 video clips collected from 40 participants, annotated for 16 types of [micro-gestures](#) that occur naturally when people are stressed or calm. These include involuntary movements like lip biting, finger tapping, or subtle shrugs, which are believed to correlate with internal stress.

The In-the-Wild Micro-Gesture Understanding and Emotion Analysis (iMiGUE) dataset is the first publicly available [micro-gestures](#) dataset designed to capture hidden or suppressed emotions. It consists of 359 lengthy videos (totalling over 2000 minutes) featuring athletes and other subjects, from which 18,499 [micro-gestures](#) samples are annotated across 32 categories. These micro-actions were captured in real-world, high-stakes environments, where individuals might display brief micro-expressions or gestures revealing genuine emotions despite maintaining an external calm demeanour.

While these datasets are valuable, none are specifically designed to focus on gestures related to "confidence" or "lack of confidence," particularly in contexts such as public speaking or video presentations. The micro-gestures associated with confidence, such as open versus closed postures and steady versus fidgety hands, may overlap with some categories found in SMG or iMiGUE. However, they are not explicitly labelled for perceived confidence. This highlights a significant data gap.

For this thesis, this gap serves as the motivation to develop a new dataset. In summary, while current micro-gesture datasets provide a starting point by focusing on stress or concealed emotions, a comprehensive resource specifically for confidence-related gestures remains lacking. This motivates the creation of a new dataset tailored to confidence-related gestures in public speaking contexts, as described in the following section.

4.1.2 CONFIDANT Dataset Development

The creation of our dataset was inspired by the previously mentioned work of Liu *et al.* [32], who focused on capturing micro-gestures from athletes during press conferences. Their dataset emphasized spontaneous gestures in high-stress situations specific to athletic contexts. However, to align with our goal of enhancing confidence in video presentations and public speaking, we shifted the domain focus from athletic press conferences to video presentations and public speaking. We selected the YouTube channel CUT [15] as the primary source for our dataset due to its rich collection of conversational upper-body videos, which are ideal for analyzing subtle human movements and micro-gestures in natural settings.

Additionally, we expanded and revised the original labelling system in the work of Liu et al. to better suit the nuances of confidence and unconfidence in public speaking contexts. Specifically, we added more labels of micro-gestures that indicate confidence and uncertainty during video presentations. This expansion addresses the unique gestures commonly observed in presenters, such as specific hand movements, posture adjustments, and facial expressions that differ from those in athletic settings.

We created 38 distinct categories, which significantly improves the dataset’s granularity for our needs (see Figure 4.1 and Table 4.1). This enhanced dataset supports robust analysis of unconfident versus confident behaviours in video presentations.

Table 4.1: Dataset Statistics and Gesture Label Counts

Statistic	Value
Total Action Categories	38
Average Videos per Action Category	10
Total Annotated Samples	412
Average Frames per Video	68

4.1.3 Classification into Confident vs Unconfident

To enable the **CONFIDANT** model to detect unconfident gestures and convert them into confident counterparts accurately, it was essential to categorize the micro-gestures in our dataset as either "Confident" or "Unconfident" categories. To achieve this, we leveraged ChatGPT [36], a large language model developed by OpenAI, to assist in the classification process. ChatGPT has been developed using extensive collections of general knowledge and human interaction patterns, allowing it to excel at interpreting body language and gestures.

We utilized a prompt designed to produce from ChatGPT the classification of each micro-gesture. The prompt used was:

I have a list of micro-gestures observed during public speaking and presentations. For each action, please classify it as either a "Confident Gesture" or an "Unconfident Gesture" based on general body language interpretations.

For more details on the classification provided by ChatGPT, refer to the Appendix B. While ChatGPT provided a useful starting point for defining confidence categories, it

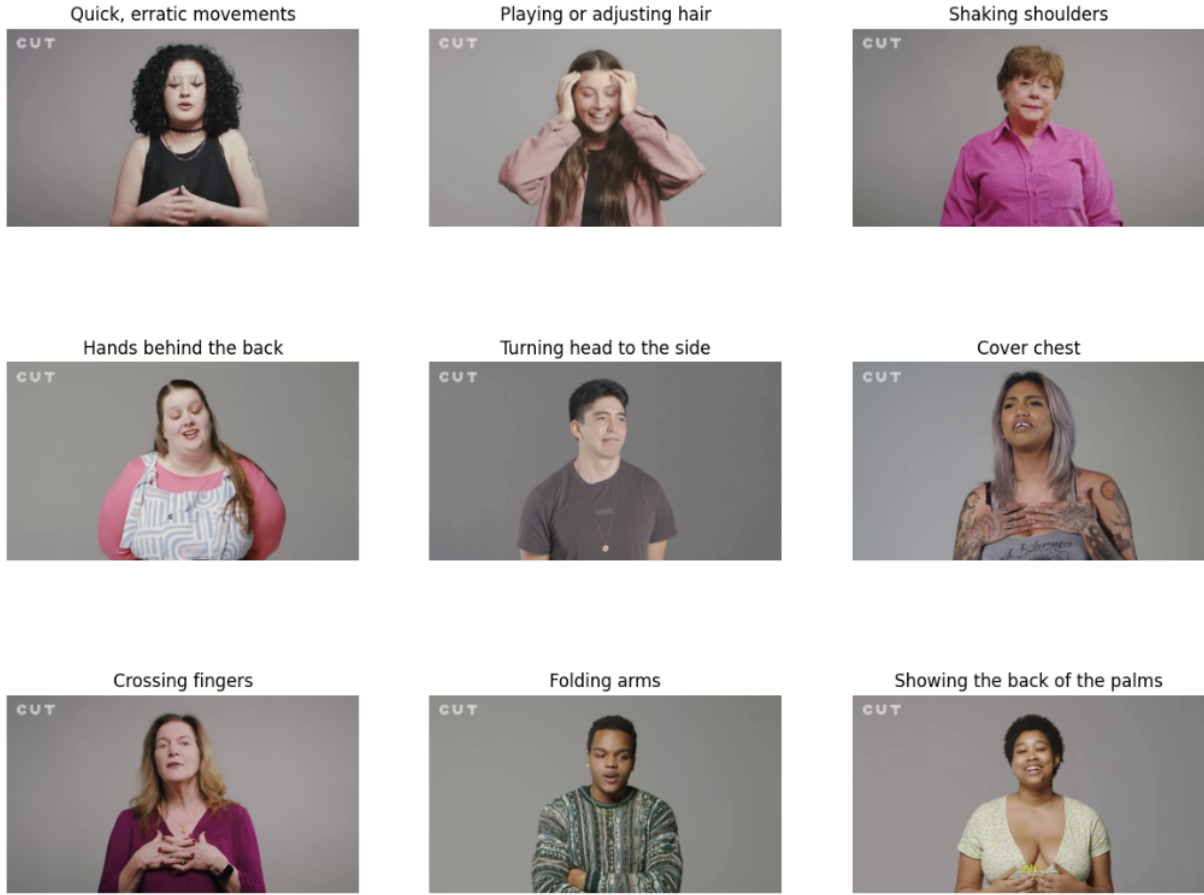


Figure 4.1: Examples of micro-gestures from our dataset.

cannot directly analyze raw video or produce frame-level annotations. A dedicated dataset and model are therefore still required to achieve reproducible, automated recognition in our pipeline. A formal dataset note is provided in Appendix C.

4.2 Micro-Gesture Action Recognition Model

The core component of [CONFIDANT](#) is a pose-based [skeleton-based action recognition](#) model that detects and classifies subtle upper-body actions in presenter-style videos. This model forms the upstream stage in the overall pipeline, providing robust and fine-grained gesture labels that are subsequently aggregated into confident/unconfident segments for

the downstream editing stage.

4.2.1 Problem Formulation

Let X denote a video segment containing T consecutive frames. From each frame $t \in \{1, \dots, T\}$, we extract a 2-D [human pose estimation](#) skeleton S_t consisting of K [keypoints](#) with spatial coordinates and confidence scores:

$$S_t = \{(x_k, y_k, c_k) \mid k = 1, \dots, K\}$$

where (x_k, y_k) are the normalized joint coordinates and $c_k \in [0, 1]$ is the [confidence score](#) from the pose estimator. The recognition problem is to learn a mapping

$$f_\theta : \{S_t\}_{t=1}^T \rightarrow y$$

where $y \in \mathcal{C}$ is one of the 38 [micro-gestures](#) categories; for confidence analysis, these are also grouped into $\mathcal{C}_{\text{confident}}$ and $\mathcal{C}_{\text{unconfident}}$.

4.2.2 Pose Representation and Keypoint Design

We utilize [DWPose](#) [56] for high-resolution whole-body [human pose estimation](#). We selected [DWPose](#) because it provides state-of-the-art accuracy for fine-grained landmarks such as hands and facial features, which are critical for distinguishing subtle [micro-gestures](#). Its two-stage distillation framework also improves robustness under varied lighting and partial occlusion, making it well-suited to presenter-style footage. Since our [near-field scenario](#) dataset predominantly consists of upper-body images, we restrict the skeleton to joints pertinent to [micro-gestures](#) recognition. Lower-body joints are omitted, as they are either absent from the frame or contribute little to confidence-related cues.

Two skeleton configurations were evaluated:

1. **14 keypoints:** Upper-body only, omitting the four leg joints from the [COCO](#) keypoint set.
2. **COCO-WholeBody (123 keypoints):** Upper-body plus comprehensive hand and facial landmarks, providing fine-grained detail for subtle gestures.

For modelling, we adopt [STGCN++](#) [17], a state-of-the-art [GCN](#) designed to capture both intra-frame spatial dependencies and inter-frame temporal dynamics. We chose [STGCN++](#) because it extends the spatio-temporal GCN family with multi-stream and adaptive weighting mechanisms, achieving strong benchmark results on skeleton-based action datasets. These capabilities make it particularly effective for modelling the subtle temporal patterns and joint interactions that characterize confidence-related micro-gestures.

4.2.3 Model Architecture

The network operates in a two-stream setup. The joint stream encodes absolute [keypoints](#), while the bone stream encodes relative displacements between anatomically connected [keypoints](#). Their fused features are passed to a classifier predicting either the 38 fine-grained [micro-gestures](#) or the binary confidence classes.

4.2.4 Experimental Setup

The model is trained on the [CONFIDANT](#) dataset. To ensure a robust model evaluation, we partitioned the dataset into training and validation sets using a 70%–30% stratified split, maintaining an approximately proportional representation of each action class. We adopt the mmAction2 toolbox [13] for training and evaluation. The training process leverages the [STGCN++](#) architecture with cross-entropy loss. Experiments were conducted on an NVIDIA RTX 3090 Ti GPU. The specific hyperparameters are listed in Table 4.2.

Hyperparameter	Value
Learning Rate	0.1
Weight Decay	5×10^{-4}
Batch Size	16
Total Epochs	40
Optimizer	SGD

Table 4.2: Training setup for our [STGCN++](#) experiments.

4.2.5 Experimental Evaluation

We evaluated the [CONFIDANT](#) micro-gesture recognition model on the [CONFIDANT](#) dataset using the [STGCN++](#) backbone, comparing two skeleton configurations: a reduced

14-keypoint upper-body set (excluding the four leg joints) and a COCO-wholebody configuration with all leg joints removed but retaining full-resolution facial and hand landmarks. This comparison was designed to quantify the impact of fine-grained hand and facial details on recognizing subtle confidence-related gestures.

In addition to the 38-class fine-grained recognition task, we also evaluated the model on the binary confident/unconfident classification task, which is directly relevant to the downstream generative editing stage. Table 4.3 reports the top-1 classification accuracy for each configuration. In the fine-grained 38-class setting, the 14-keypoint configuration achieved a top-1 accuracy of 47.3%. Switching to the COCO-wholebody (legs excluded) configuration increased performance to 67.0%, representing a substantial +19.7% absolute improvement.

This improvement demonstrates that high-resolution hand landmarks and facial keypoints capture discriminative motion patterns, such as finger fidgeting, clasped hands, or subtle posture shifts, that are not well represented in the reduced skeleton. In the binary classification setting (*confident* vs. *unconfident*), the model achieved 87.0% top-1 accuracy, demonstrating its robustness in reliably segmenting video into confidence-based intervals for downstream processing. These results confirm that both the selection of keypoints and the inclusion of detailed hand and facial landmarks are critical for accurate micro-gesture recognition.

Table 4.3: Top-1 accuracy on the **CONFIDANT** dataset using **STGCN++** for different skeleton configurations. Results are reported for the 38-class fine-grained classification task and the binary confident/unconfident classification task.

Skeleton Configuration	Top-1 Accuracy (%)
14 keypoints (upper-body only)	47.3
COCO-wholebody (upper-body only)	<u>67.0</u>
Binary (confident vs. unconfident)	87.0

To complement these quantitative findings, Figure 4.2 presents qualitative examples for the binary classification task. All shown predictions are correct, illustrating the model’s ability to generalize across individuals, clothing styles, and subtle variations in posture or hand movement. In the confident class, the model successfully recognizes open and expansive postures such as arms akimbo (Figure 4.2a) and standing tall with an upright posture (Figure 4.2c). For the unconfident class, it correctly identifies more closed or self-directed motions such as fidgeting fingers (Figure 4.2b) and rubbing eyes (Figure 4.2d).

These examples demonstrate the recognizer’s ability to detect subtle gestures, even when the differences between classes involve fine-grained hand or facial cues rather than large-scale body movements. Taken together, the results in Table 4.3 and Figure 4.2 demonstrate that the **CONFIDANT** model can deliver both accurate fine-grained micro-gesture recognition and reliable binary classification of confident versus unconfident behaviour. This performance is critical for the downstream editing pipeline, as it ensures that transformations are applied selectively and only to those segments that are truly unconfident, while preserving confident gestures unchanged.

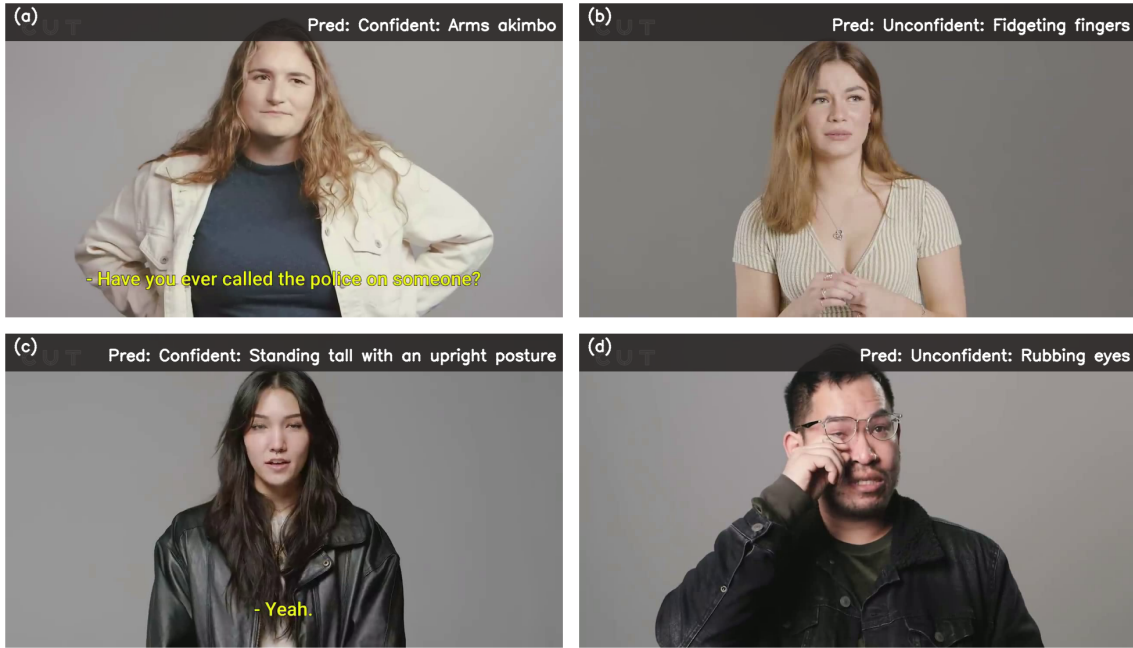


Figure 4.2: Qualitative results for the binary confident/unconfident classification task using **STGCN++** with the COCO-wholebody (legs excluded) configuration. Each sub-figure shows a correctly classified example: (a) confident—arms akimbo; (b) unconfident—fidgeting fingers; (c) confident—standing tall with an upright posture; (d) unconfident—rubbing eyes. These examples illustrate the model’s robustness across diverse individuals, gesture types, and visual conditions.

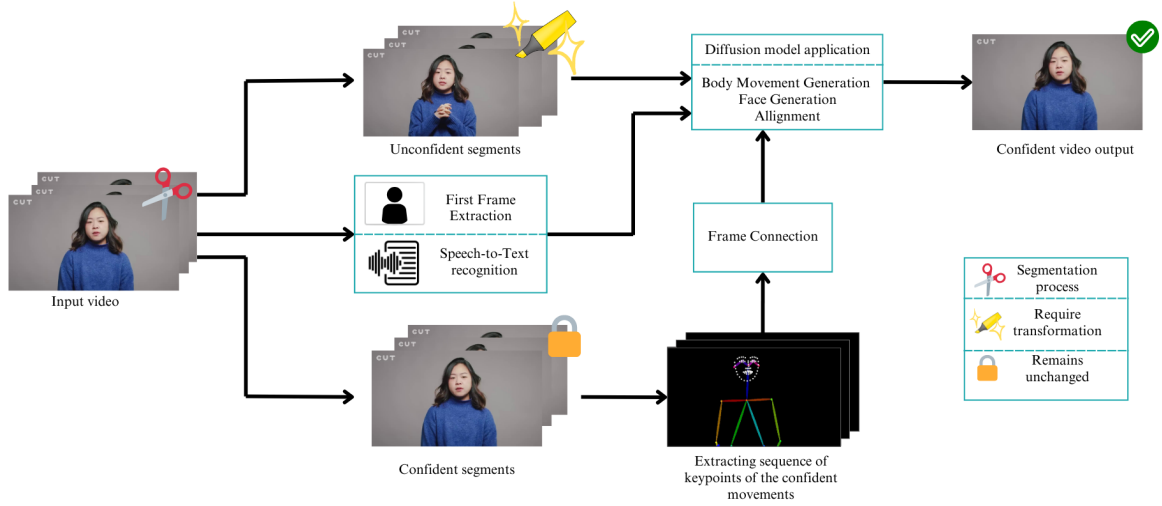


Figure 4.3: **Overview of the CONFIDANT Generative Video Editing Pipeline.** The input video is divided into confident and unconfident segments. **keypoints** from confident parts are used to model confident gestures, with some serving as references for transformation and others for interpolating joints. These **keypoints** and reference frame are processed through a diffusion model to generate confident body movements. A second model then creates facial expressions and synchronizes lip movements, utilizing a zoomed-in face and altered audio. The outputs are aligned to produce a seamless transformed segment that blends with untouched parts.

4.3 Application: Identity-Preserving Confidence Editing (Downstream Task)

In designing CONFIDANT, our primary objective was to enhance perceived confidence in video presentations by transforming unconfident **micro-gestures** into confident ones while maintaining the speaker’s identity. This involves addressing both visual and auditory elements through four interconnected components: video segmentation and speech alignment (Section 4.3.1), **keypoint** extraction (Section 4.3.1), diffusion-based confident gesture editing (Section 4.3.2), and facial expression and lip movement generation (Section 4.3.3). These components are illustrated in Figure 4.3.

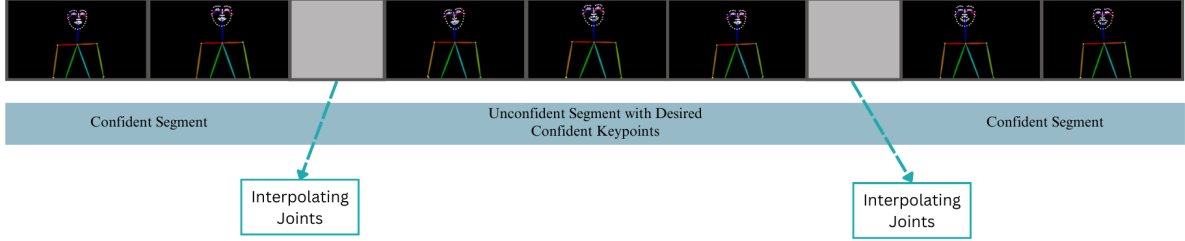


Figure 4.4: **Frame Connection.** Keypoints from the confident segments and the modified unconfident segments are connected using joint interpolation to create a continuous and natural transition across frames, resulting in fluid and consistent gestures.

4.3.1 Preprocessing

Video Segmentation and Speech Alignment

Segmentation Process: We segment the input video into unconfident and confident parts using the binary classifier described in Section 4.2. This allows us to modify unconfident behaviours while leaving confident segments untouched.

Speech Alignment and Cleanup: To ensure smooth transitions and maintain coherence in the edited video, we perform speech alignment by synchronizing audio with video frames. Additionally, we enhance delivery by identifying and removing filler words and disfluencies.

Keypoint Extraction

For each frame in the confident segments, we extract body keypoints using DWPose [56]. The model detects and tracks keypoints (e.g., shoulders, elbows, wrists). The extracted keypoints provides a skeletal representation of posture and movements, which is essential for understanding and modifying gestures.

4.3.2 Diffusion Model for Confident Gesture Editing

Transforming unconfident micro-gestures into confident ones requires generating new body movements that are natural and seamlessly integrated with the original video. To achieve this, we employ a diffusion-based generative model that leverages the extracted keypoints to produce confident gestures.

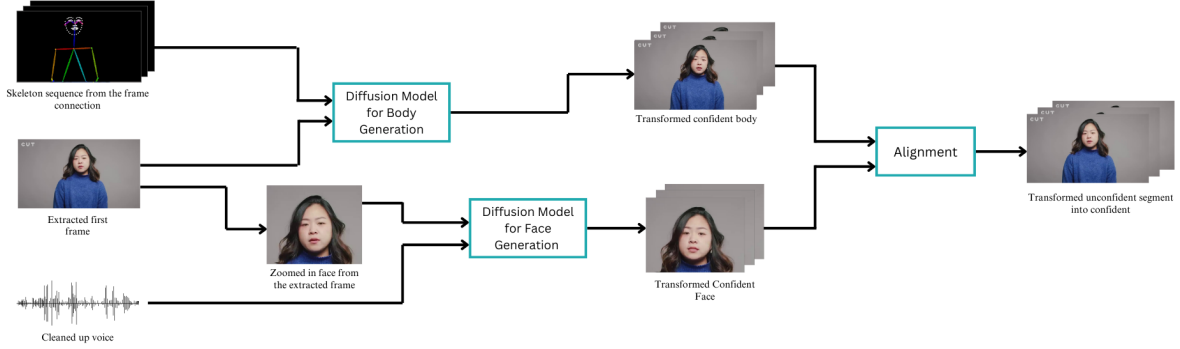


Figure 4.5: The process begins with **Body Movement Generation**, where the reference frame and skeleton **keypoints** are input into a diffusion model to create transformed body movements. Next, **Facial Expression and Lip Movement Generation** uses a zoomed-in face and cleaned audio in a facial diffusion model to produce synchronized lip movements. Finally, in the **Alignment Process**, these outputs are aligned to form a cohesive, confident segment.

After **keypoint** extraction, two critical components are involved: *Frame Connection* and *Body Movement Generation*.

Frame Connection: To generate smooth transitions, we interpolate **keypoints** across frames preceding and following the transformed segment, creating continuous confident gestures (Figure 4.4).

Body Movement Generation: We employ Moore-AnimateAnyone [43], an open-source diffusion model chosen for its relatively low computational cost compared to other state-of-the-art motion generation frameworks. It synthesizes body movements that match the target **keypoints**, taking the target **keypoints** sequence and the initial frame as input to generate the confident video segment (Figure 4.5).

4.3.3 Diffusion Model for Facial Expression and Lip Movement Generation

The next step is to ensure facial expressions and lip movements are natural and synchronized with the audio.

Face and Lip Syncing: While Moore-AnimateAnyone generates body movements, it

does not synchronize lips with audio. To address this, we use Hallo [53], which leverages aligned and cleaned audio to produce realistic lip movements (Figure 4.5).

Alignment: Post-processing aligns generated facial expressions with body movements. Spatial alignment ensures that facial features are positioned correctly; temporal smoothing and blending harmonize both modalities, resulting in a natural and confident output.

4.3.4 Experimental Setup

We evaluated CONFIDANT’s effectiveness by transforming unconfident segments and comparing them with original confident segments. Experiments were conducted on an NVIDIA RTX 3090 Ti GPU. We utilized Moore-AnimateAnyone [43] for body movements and Hallo [53] for lip synchronization. The diffusion models processed videos at 896×512 resolution over sequences of 64 frames.

4.3.5 Quantitative Results

We evaluated with standard metrics used in video generation:

- **FVD** [47]: Measures similarity between real and generated videos, using action-recognition features to capture spatio-temporal consistency. Lower is better.
- **SSIM** [50]: Compares luminance, contrast, and structural information. Ranges 0–1; higher is better.
- **PSNR** [1]: Ratio between maximum signal power and noise power. Higher is better.
- **LPIPS** [58]: Perceptual similarity using deep features. Lower is better.

Table 4.4 summarizes the evaluation metrics for the best-performing model. The low FVD score of 507.74 indicates that the transformed videos closely resemble real videos in terms of motion dynamics and visual appearance, suggesting effective maintenance of natural patterns. Additionally, high SSIM (0.9082) and PSNR (25.19 dB) values indicate that the structural and pixel-wise similarities between the original and transformed videos are well-preserved, suggesting minimal distortion during the transformation. Finally, the low LPIPS value of 0.0225 reflects high perceptual similarity, confirming that the transformed videos appear natural and consistent to viewers.

Table 4.4: Quantitative Evaluation Metrics

Metric	Value
FVD	507.74
SSIM	0.9082
PSNR	25.19 dB
LPIPS	0.0225

4.3.6 Qualitative Results

To qualitatively assess the effectiveness of the [CONFIDANT](#) model, we focused on transforming unconfident segments within the same input video illustrated in Figure 4.6. We began by extracting the first frame and the facial features from the original image to serve as reference images for preserving the individual’s identity. The detected unconfident segment was then transformed using [CONFIDANT](#) to generate a confident version of the same segment. This approach demonstrates the model’s ability to enhance confidence in gestures while preserving the person’s appearance and natural demeanour.

4.3.7 Ethical Statement and Risks of Misuse

This chapter explores identity-preserving generative editing of micro-gestures to alter perceived "confidence". Such editing creates clear risks. Edited clips can mislead viewers about a person’s intent or demeanour and be used manipulatively; applying edits without the subject’s consent violates autonomy; and inferring latent traits from visible behaviour risks stigma and discrimination. Perceptions of "confidence" vary across culture, context, disability, and neurodiversity, so the binary mapping {Confident, Unconfident} is heuristic rather than ground truth. Any deployment in hiring, credit, insurance, political persuasion, surveillance, or reputational contexts is inappropriate and outside the scope of this research. To reduce harm, we treat the system and the results as research-only. We do not redistribute raw third-party video; releases, if any, are limited to code and derived artifacts (labels, keypoints, IDs) according to platform terms (Appendix C).

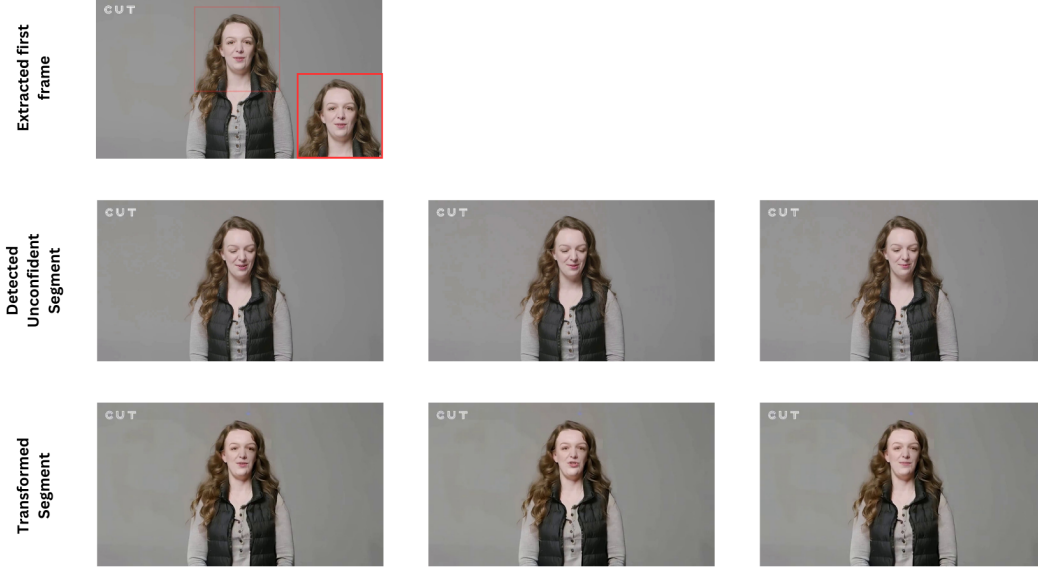


Figure 4.6: Qualitative comparison. First row: Extracted first frame and facial reference images; Second row: Detected unconfident segment, showcasing gestures of low confidence; Third row: Transformed confident segment produced by [CONFIDANT](#), exhibiting enhanced gestures conveying confidence.

4.4 Conclusion

This chapter introduced the [CONFIDANT](#) framework for [near-field scenario](#), fine-grained recognition of confidence-related [micro-gestures](#) and its application to identity-preserving generative video editing. To address the absence of suitable resources, we curated a dataset of 412 annotated clips across 38 [micro-gestures](#) categories, sourced from presenter-style videos. Unlike existing datasets such as SMG and iMiGUE, the [CONFIDANT](#) dataset explicitly labels gestures by perceived confidence, enabling both binary classification and fine-grained recognition.

We developed and evaluated a pose-based [skeleton-based action recognition](#) model using [STGCN++](#). Two skeleton configurations were tested: a 14-[keypoint](#) upper-body set and a [COCO-WholeBody](#) (legs excluded) with detailed hand and face landmarks. Results showed a gain from 47.3% to 67.0% [Top-1](#) accuracy in the 38-class task, highlighting

the importance of fine-grained landmarks. In the binary task, accuracy reached 87.0%, confirming the model’s reliability for downstream segmentation. Qualitative examples confirmed the correct identification of both confident and unconfident gestures.

The recognizer forms the upstream stage of the **CONFIDANT** pipeline, providing temporal boundaries for editing. In the downstream application, a diffusion-based framework selectively transforms unconfident gestures into confident ones while preserving identity, synchronization, and unaffected confident segments. Quantitative metrics (**FVD**=507.74, **SSIM**=0.9082, **PSNR**=25.19 dB, **LPIPS**=0.0225) confirm high visual and perceptual fidelity, while qualitative examples demonstrate seamless integration.

Overall, this chapter establishes a comprehensive **near-field scenario** pipeline, spanning dataset curation, high-precision recognition, and targeted generative editing. By combining fine-grained **human pose estimation** with synthesis models, **CONFIDANT** advances the field of **micro-gestures** understanding in presenter-style videos, paving the way for real-time feedback systems in communication skills training.

Chapter 5

ZoomGate: Scale-Aware Action Recognition Across Mixed Zoom Levels

Cinematic video rarely conforms to a single viewpoint: a scene may shift from a wide establishing shot to a medium dialogue frame to an extreme close-up within seconds, each emphasizing different cues. This variability poses a scale-variance problem for action recognition, since the discriminative evidence depends on whether the face, torso, hands, or full body are visible. To address this, we propose [ZoomGate](#). This [scale-aware routing](#) framework first assigns each person instance to a zoom-view category and then routes the segment to specialized recognition models. This decomposition enables each recognizer to operate within its domain of strength, thereby improving robustness across various mixed cinematic conditions.

The remainder of this chapter is organized as follows. Section [5.1](#) defines the problem setting and motivates the need for scale-aware reasoning in cinematic video. Section [5.2](#) introduces the semi-supervised dataset of movie trailers used to evaluate [ZoomGate](#). Section [5.3](#) presents the overall pipeline design, while Section [5.4](#) details the baseline heuristic rules, analyzes systematic failure cases, and describes our refined rules with temporal smoothing, supported by qualitative examples. Section [5.5](#) evaluates downstream action recognition performance across zoom levels, using specialized recognizers for face, hand, [micro-gestures](#), and full-body views. Finally, Section [5.6](#) concludes with a discussion of [ZoomGate](#) as a unifying bridge between [near-field scenario](#) and [far-field scenario](#) recognition in cinematic video.

5.1 Problem Definition & Motivation

Action recognition in cinematic video poses unique challenges that extend beyond those observed in either [near-field scenario](#) (presenter videos) or [far-field scenario](#) (broadcast sports) domains. Unlike these relatively homogeneous settings, movies and trailers are characterized by rapid alternation between camera scales: a wide establishing shot may cut to a medium shot of two characters, followed by an extreme close-up face within a matter of seconds. Each shot type emphasizes a different subset of visual cues: whole-body motion in long shots, hand gestures or torso posture in medium shots, and micro-expressions in close-ups. For a recognition system, this introduces a critical scale-variance problem: the features that are most discriminative for one view may be absent in another.

Traditional [HAR](#) pipelines typically employ a one-size-fits-all approach, using the same model architecture regardless of the camera scale. While effective in controlled benchmarks, this approach is fragile in real-world cinematic material where a close-up can eliminate body pose cues, or a wide shot can render faces too small for expression analysis. As a result, such models either fail to consistently recognize actions across shots or overfit to the most frequent shot scale in the dataset.

We therefore define the problem as follows: given raw movie footage with arbitrary camera zooms and cuts, assign each [tracklet](#) to an appropriate zoom-view category ([FACECLOSE](#), [HANDSONLY](#), [UPPERBODY](#), [FULLBODY](#), or [NOPARTS](#)) and route the resulting temporally stable segments to scale-appropriate action recognition models. The core difficulty lies in making this view assignment robust under detector noise, occlusions, and rapid transitions, while ensuring that the downstream recognizers receive inputs with maximally relevant cues.

The motivation for [ZoomGate](#) is twofold. First, by explicitly acknowledging the heterogeneity of camera views in movies, we can leverage specialized recognizers: fine-grained gesture models in close-up segments, and whole-body activity models in long shots. This division of labour allows each model to operate in its domain of strength, improving overall robustness. Second, the view assignment itself provides a form of interpretable intermediate representation. By exposing when the system “believes” it is in a close-up versus a long shot, we gain diagnostic visibility into failure cases and a principled way to align recognition with cinematographic conventions. In this sense, [ZoomGate](#) is not only a unification of [near-field scenario](#) and [far-field scenario](#) recognition strategies, but also a bridge to the complex, [zoom-mixed scenario](#) reality of cinematic video.

5.2 Semi-Supervised Movie Dataset

To evaluate our heuristic zoom-level labelling and temporal smoothing pipeline, we constructed a semi-supervised dataset from publicly available movie trailers. In total, eight trailers were selected, encompassing a diverse range of genres, including drama, action, comedy, and animation. The trailer durations ranged from 1 to 5 minutes, resulting in a dataset with heterogeneous visual styles and cinematographic conventions. This diversity was essential to test the robustness of our rules across both live-action and animated material. A resource note summarizing components, inputs/outputs, and availability is given in Appendix D.

5.2.1 Semi-Supervised Annotation Process

The annotation process followed a semi-supervised approach. First, we applied automatic detection using the YOLOv8 model [48] for person BBs, together with MediaPipe keypoint estimators [33] for face, hands, and body keypoints. These raw detections provided the foundation for determining per-frame spatial configurations. Based on these detections, we introduced a set of heuristic rules (see Section 5.4) to assign a zoom-level label to each frame: FACECLOSE, HANDSONLY, FULLBODY, UPPERBODY, or NOPARTS. This strategy allowed us to scale annotation across long videos without requiring manual frame-by-frame labelling.

5.2.2 Dataset Statistics

Table 5.1 summarizes the dataset composition. The label distribution highlights that the dataset is naturally imbalanced, with UPPERBODY views dominating and HANDSONLY being notably underrepresented due to the prevalence of dialogue-driven scenes.

Table 5.1: Statistics of the semi-supervised movie dataset.

Videos	Frames	Zoom Labels Distribution
8 trailers	28,880	FACECLOSE: 4.36%
		HANDSONLY: 0.62%
		FULLBODY: 12.66%
		UPPERBODY: 60.18%
		NOPARTS: 22.18%

5.2.3 Manual Dataset Verification

In addition to automatic labelling, we also performed manual verification using a custom-built graphical interface (Figure 5.1). This tool enabled us to scroll through video frames, inspect BBs and labels, and verify whether the assigned zoom level was suitable. By integrating both automatic and human-in-the-loop inspection, we ensured higher reliability of the semi-supervised annotations.

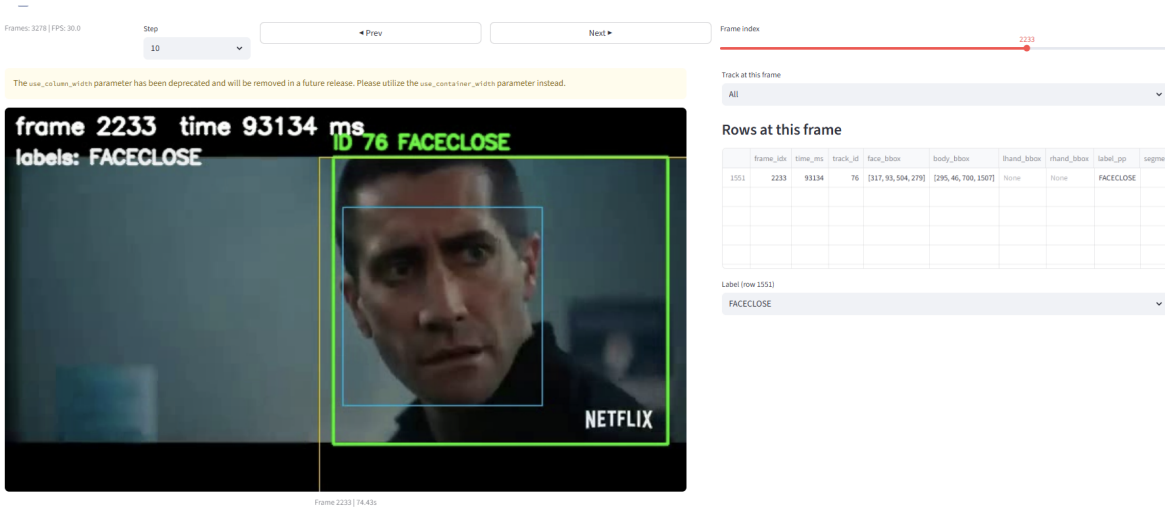


Figure 5.1: Screenshot of the manual verification interface. The GUI enables frame navigation and inspection of automatically assigned labels, providing a human-in-the-loop mechanism for correcting edge cases.

5.3 Pipeline Overview

The [ZoomGate](#) framework is organized as a sequence of interconnected stages, each addressing a specific subproblem in the recognition pipeline. This modular decomposition separates the task of scale estimation (“which visual cues are visible in the frame?”) from action classification (“what action is being performed?”), which improves robustness to cinematic variation. By routing each segment to a scale-appropriate recognizer, the framework leverages the strengths of both [near-field scenario](#) and [far-field scenario](#) models. An overview of the complete pipeline is shown in Figure 5.2.

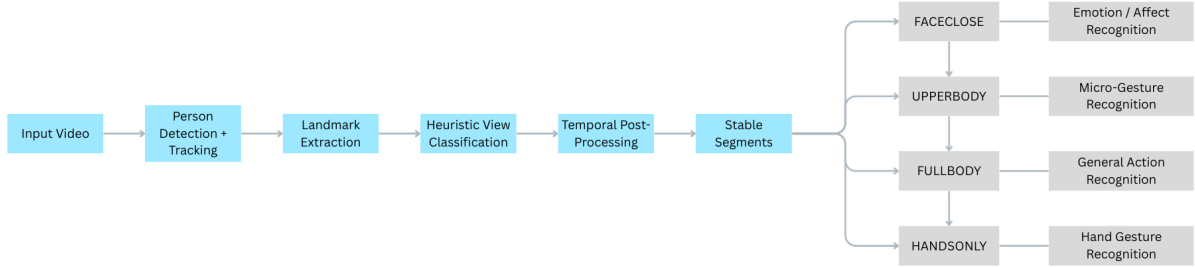


Figure 5.2: Overview of the [ZoomGate](#) pipeline. Raw video is processed into person tracks, assigned to zoom-view categories, and routed to specialized recognition models: emotion recognition for close-ups, [micro-gestures](#) recognition for upper-body views, general [HAR](#) for full-body views, and hand gesture recognition for hand-only shots.

As illustrated in the figure, the pipeline consists of five main stages: (1) person detection and tracking, (2) [keypoint](#) extraction, (3) heuristic view classification, (4) temporal post-processing, and (5) action recognition routing. Each stage is described in detail below.

5.3.1 Stage 1: Person Detection and Tracking

The input video is first decomposed into frame-level detections using YOLOv8 [48], chosen for its strong balance of speed and accuracy in person detection. To preserve temporal consistency across frames, detections are linked using the BoT-SORT tracker [3], a lightweight and reliable tracker that maintains stable identities under fast motion and occlusion. This ensures that all downstream operations operate on coherent [tracklets](#) rather than on single-frame [BBs](#).

5.3.2 Stage 2: Keypoint Extraction

For each tracked individual, we extract fine-grained anatomical cues with MediaPipe [33], including facial [BBs](#), full-body pose [keypoints](#), and hand [keypoints](#). These cues provide the necessary evidence for distinguishing different zoom levels and are more robust indicators than [BBs](#) alone.

5.3.3 Stage 3: Heuristic View Classification

Based on the extracted cues, a heuristic decision function (see Section 5.4) assigns one of five zoom-view labels (FACECLOSE, HANDSONLY, UPPERBODY, FULLBODY, or NOPARTS) to each frame–track pair. The rules integrate facial scale, torso and limb visibility, hand plausibility, and occlusion guards to provide a robust mapping under diverse cinematic conditions.

5.3.4 Stage 4: Temporal Post-Processing

Per-frame labels are inherently noisy due to detector jitter and tracker switches. To stabilize the predictions, we apply a temporal smoothing algorithm (see Section 5.4.4), which merges short runs, bridges small gaps, and outputs temporally consistent view labels. Each stable run of at least 30 frames (1 second at 30 [fps](#)) is exported as a coherent segment.

5.3.5 Stage 5: Action Recognition Routing

Finally, the stable segments are routed to recognition models specialized for their respective view classes. For example:

- FACECLOSE → emotion / affect recognition models,
- HANDSONLY → hand gesture recognizers,
- UPPERBODY → [micro-gestures](#) recognizers,
- FULLBODY → general [HAR](#) models.

Segments labelled as NOPARTS are filtered out, since they correspond to frames without discernible human parts, such as occluders or background shots, and therefore provide no meaningful input for action recognition.

This modular pipeline enables [ZoomGate](#) to account for scale heterogeneity explicitly, producing interpretable zoom-view labels and improving recognition performance across [zoom-mixed scenarios](#) cinematic conditions.



Figure 5.3: Illustrative examples of the five zoom-view categories used in [ZoomGate](#): (a) UPPERBODY, (b) HANDSONLY, (c) NOPARTS, (d) FACECLOSE, and (e) FULLBODY. These examples, drawn from cinematic footage, provide intuitive reference points before introducing the heuristic decision rules.

5.4 Heuristic Rules for View Classification

A core component of the [ZoomGate](#) pipeline is the assignment of each person-frame instance to one of five zoom-view categories (see Figure 5.3): FACECLOSE, HANDSONLY, UPPERBODY, FULLBODY, or NOPARTS. Our baseline formulation relied on simple geometric thresholds, such as the ratio of the detected face area to the frame size or the presence of lower-limb [keypoints](#). While effective in controlled cases, these rules proved brittle under cinematic conditions, which often involve rapid cuts, foreground occlusions, and extreme variation in [contextual information](#) scale.

5.4.1 Baseline Rules

The initial decision criteria are summarized in Table 5.2. For example, a frame was classified as FACECLOSE if the detected face occupied more than 7% of the image area or exceeded 30% of the image height, and as FULLBODY if the knees or ankles were visible. Frames with no visible [keypoints](#) were mapped to NOPARTS.

Table 5.2: Baseline heuristic rules with brief descriptions of each zoom-view label.

Label	Baseline Criteria	Description / Typical Shot
FACECLOSE	Face area $> 7\%$ of frame or face height $> 30\%$ of frame height	Extreme close-up/tight head shot; facial micro-expressions dominate; little or no torso visible.
HANDSONLY	Hands detected, but no face or body	Cropped arms/hands region; object-centric or gesture emphasis shots.
FULLBODY	Pose detected with lower-body keypoints (knees, ankles)	Long/wide shot; entire silhouette visible; whole-body motion cues available.
UPPERBODY	Body detected but not full body (torso up only)	Medium / medium close-up; head + shoulders/torso; arm/torso cues present, legs absent.
NOPARTS	No face, hands, or body keypoints detected	No valid human parts (e.g., occluders, empty frames, non-human content); excluded from recognition.

5.4.2 Failure Cases and Edge Conditions

During exploratory labelling, several systematic edge cases emerged that motivated refinement of the rules. These issues stemmed from tracker ID instability, occlusions, scale ambiguity, and overly sensitive [keypoint](#) detections. Representative examples are shown in Figure 5.4.

5.4.3 Updated Rules

These systematic errors motivated the design of updated rules that incorporate part-based evidence, occlusion guards, and scale priors. The full pseudocode implementation is provided in Appendix E for reproducibility. Table 5.3 summarizes the final set of rules.



Figure 5.4: Representative failure cases that motivated updates to the heuristic rules. (a) Tracker ID switching across consecutive frames, leading to inconsistent view assignments. (b) Far-scale character misclassified as `UPPERBODY` instead of `FULLBODY`. (c) Foreground occluder causes background torso to be missed and misclassified as `NOPARTS`. (d) Over-sensitive knee detection results in unstable toggling between `UPPERBODY` and `FULLBODY`.

Key Principles.

1. **Evidence gate.** A body view (`UPPERBODY` or `FULLBODY`) is only assigned if at least one reliable cue is present: a valid face, at least two torso joints, or consistent hand detections.
2. **Hands-only validity.** A hand is only considered valid if its relative size and vertical position are plausible with respect to the body `BB` (top 70% of body height and 2–25% of body area).
3. **Face vs. upper body disambiguation.** A large detected face normally indicates `FACECLOSE`, but if at least two torso joints are also visible, the label is upgraded to `UPPERBODY`.
4. **Far-person fallback.** If the body `BB` is very small ($<30\%$ of frame height) and the face is correspondingly small ($<12\%$ of frame height), the label defaults to `FULLBODY`, even if lower-limb evidence is missing.
5. **Edge-occluder guard.** Large `BBs` that touch two or more image borders but contain no part evidence are forced to `NOPARTS`.

Label	Updated Rule (ZoomGate)
FACECLOSE	Face area $> 7\%$ of frame or face height $> 30\%$ of frame height; unless ≥ 2 torso joints are visible, in which case reclassify as UPPERBODY.
HANDSONLY	Valid hand(s) detected (top 70% of body box, 2–25% of body area), with no face, torso evidence, or lower-limb visibility .
FULLBODY	Either (i) lower limbs (knees/ankles) visible and some upper evidence (face or torso), or (ii) far-person fallback: body BB height $< 30\%$ of frame and non-close face ($< 12\%$ of frame height).
UPPERBODY	Face detected or at least two torso joints visible, but without reliable lower-limb evidence. Covers medium close-ups.
NOPARTS	No valid face, torso, hands, or lower-limb cues. Also applied to large BBs touching ≥ 2 image borders (edge-occluder guard) or blurred regions (optional).

Table 5.3: Updated heuristic rules for zoom-view classification. These rules extend the baseline criteria (Table 5.2) by addressing the edge cases illustrated in Figure 5.4, through explicit checks for evidence, occlusion, and scale.

5.4.4 Temporal Smoothing

Even with refined rules, per-frame predictions can remain unstable due to tracker ID switches and **keypoint** jitter. To address this, we apply a temporal post-processing stage that operates at the **tracklet** level. First, a rolling mode filter with a seven-frame window reduces frame-to-frame flicker, ensuring that transient misclassifications do not dominate. Next, short runs of fewer than 30 frames are merged into neighbouring segments, with preference given to the longer or higher-confidence segment. Finally, small gaps of up to ten frames between longer runs of the same label are bridged, yielding more coherent temporal spans. Through these steps, noisy framewise labels are converted into stable segments of at least one second (30 frames at 30 **fps**), which can then be exported as independent clips for downstream **HAR**.

5.4.5 Qualitative Results for Zoom Labels

To evaluate the effectiveness of the updated heuristics, we present a set of qualitative examples covering a diverse range of cinematic conditions. Figure 5.5 illustrates six representative cases in which the system successfully assigned zoom-view labels that are consistent with human intuition, highlighting both the interpretability of the [ZoomGate](#) framework and the robustness of the revised rules.

In panel (a), a close-up of a helmet is correctly recognized as **NOPARTS**, since the system avoids misclassifying the object as a face. Panel (b) demonstrates that the framework generalizes beyond live-action footage: in an animated crash sequence with cluttered motion, the presence of characters is reliably identified. Panel (c) shows an isolated hand gesture cropped tightly to the arms, which is successfully mapped to the **HANDSONLY** category, illustrating the benefit of dedicated part-based rules. In panel (d), a medium close-up portrait containing both the head and part of the torso is distinguished from a pure facial close-up, leading to the correct assignment of the **UPPERBODY** label. Panel (e) presents a full-body walking sequence where a character enters through a doorway; even in the absence of high-confidence lower-limb detections, the system robustly assigns the **FULLBODY** category. Finally, panel (f) illustrates a crowded dining scene with multiple overlapping people, where the system maintains consistent labelling across several individuals simultaneously.

Together, these examples demonstrate that the updated heuristics not only resolve systematic edge cases (Section 5.4) but also yield stable, interpretable view assignments across a wide variety of genres, shot types, and visual styles.

5.5 Action Recognition

Having established stable zoom-view segments, the final stage of [ZoomGate](#) is to perform [HAR](#) tailored to each view class. This section presents the design and evaluation of this stage. We begin by describing the curated [HAR](#) dataset and its statistics, then introduce the specialized models used for each zoom-view category: hand gesture recognition for **HANDSONLY**, facial emotion recognition for **FACECLOSE**, upper-body micro-gesture recognition for **UPPERBODY**, and full-body action recognition for **FULLBODY**. For each case, we report qualitative examples that illustrate the model’s predictions and demonstrate how scale-specific recognizers contribute to a more robust and interpretable overall framework.

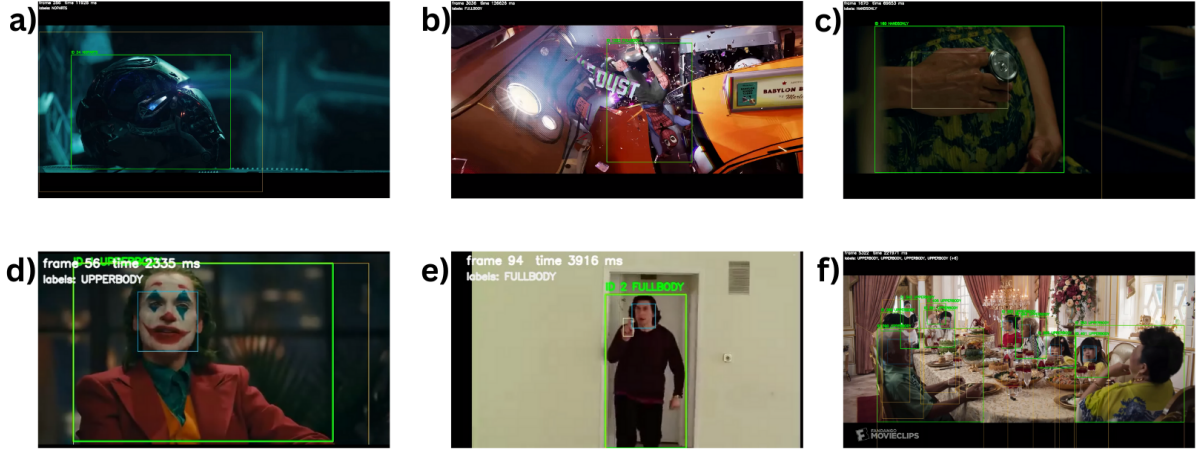


Figure 5.5: Qualitative results of zoom-view. (a) Tight helmet close-up detected as no parts, (b) animated crash sequence, (c) isolated hand shot, (d) medium close-up with torso visible, (e) full-body walking, (f) crowded multi-person dining scene. The assigned labels align well with human intuition across varied contexts.

5.5.1 Dataset and Statistics

To evaluate the action recognition stage of [ZoomGate](#), we curated a dataset of 319 clips (46,284 frames, 1,909 seconds) derived from raw video material. Each clip corresponds to a coherent action instance aligned with a zoom-view category, forming the atomic units for training and evaluation.

As shown in Table 5.4, the dataset is naturally imbalanced: **UPPERBODY** dominates due to the prevalence of dialogue-driven medium shots, while **FULLBODY** and **FACECLOSE** occur less often, and **HANDSONLY** remains rare. This imbalance reflects typical cinematic patterns and motivates the need for scale-aware routing, since recognizers must adapt to the varying availability of cues across views.

5.5.2 Hand-Only Gesture Recognition

A special case in the dataset is the **HANDSONLY** category, which focuses on recognizing isolated hand gestures. Unlike the broader body- or face-centric clips, these segments are restricted to localized hand regions, offering a more fine-grained view of manual gestures.

Label	Clips	Frames	Seconds
UPPERBODY	266	43,300	1784.5
FULLBODY	34	1,957	81.6
FACECLOSE	18	995	41.5
HANDSONLY	3	101	4.2
Total	319	46,284	1909.8

Table 5.4: Per-label statistics of the [HAR](#) dataset. The dataset excludes `NOPARTS` clips, since no action can be inferred in the absence of valid human cues. The imbalance reflects typical cinematic patterns: `UPPERBODY` dominates due to dialogue-driven medium shots, while close-ups (`FACECLOSE`), long shots (`FULLBODY`), and rare hand-centric frames (`HANDSONLY`) are less frequent.

For this task, we utilize the X-CLIP model [34], a multimodal vision-language [transformer](#) developed initially for video understanding. By encoding both visual sequences and text prompts representing possible gesture labels, X-CLIP enables zero-shot classification of hand gestures. We selected X-CLIP because it generalizes to a broad vocabulary of gestures without requiring a dedicated hand-gesture training set, which is scarce and often domain-specific. In contrast, most models trained solely on hand actions are limited to predefined label sets and require costly retraining to extend to new gestures. In practice, X-CLIP’s flexibility allows us to probe diverse hand-centric actions directly, making it a better fit for our heterogeneous cinematic material. A complete list of the final gesture label set is provided in Appendix F.

To further evaluate the performance of our [HAR](#) system, we examined qualitative results for the `HANDSONLY` predictions. Figure 5.6 presents representative frames extracted from three sequences, together with the model’s predicted actions. While the predictions are not exact matches to the ground truth, they are often semantically close: for instance, a sequence resembling a "push away" gesture was instead classified as holding or moving an object, a clasped-hands posture was predicted as a generic hand-holding gesture, and a "making a phone call" pose was interpreted as taking a picture. These near-misses illustrate both the strengths and the limitations of zero-shot recognition in this setting. The model is sensitive to local hand configurations and produces plausible interpretations, yet fine-grained disambiguation between visually similar gestures remains challenging.

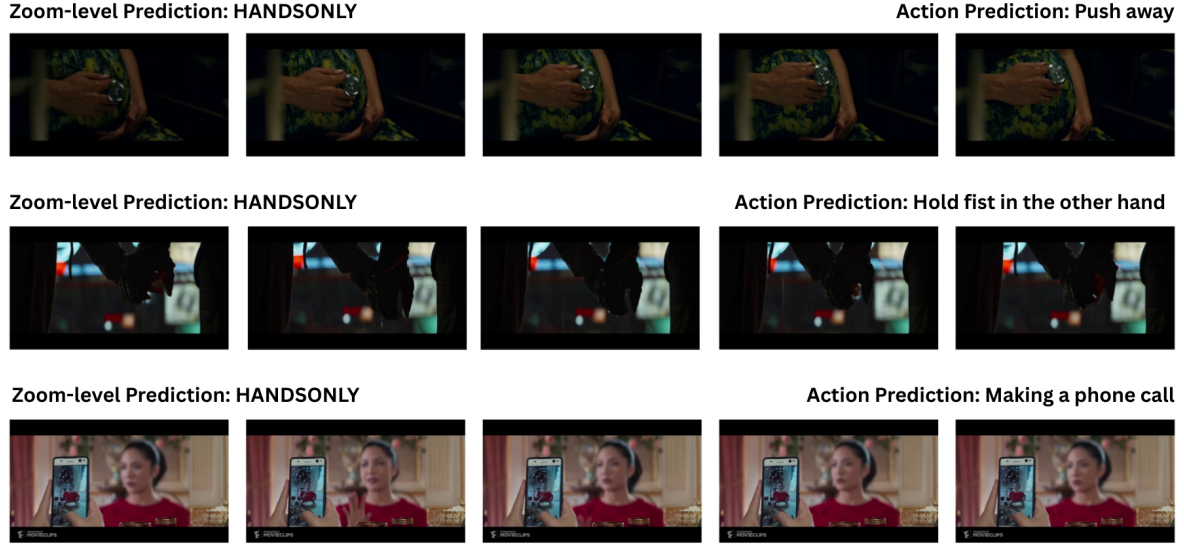


Figure 5.6: Qualitative examples of HANDSONLY predictions for hand gesture recognition.

5.5.3 Facial Emotion Recognition

To capture fine-grained affective states, we extended the pipeline to facial emotion recognition (FER). This stage focused on close-up face clips extracted from the raw videos, where emotional expressions could be reliably detected. We employed the DeepFace framework [40], which provides state-of-the-art pretrained backbones for expression analysis across multiple emotions.

For each video, we sampled a fixed number of evenly spaced frames (default: 5) and applied a range of DeepFace’s **analyze** method restricted to the emotion action space. Predictions were collected on a per-frame basis and subsequently aggregated using a majority voting strategy. The dominant emotion label per video was thus defined as the emotion most frequently predicted across sampled frames, ensuring robustness against transient misclassifications or noise.

Figure 5.7 shows representative qualitative results of the FER system. The system successfully captures a diverse range of expressions, including happiness, neutrality, sadness, and fear, even under challenging lighting and occlusion conditions. While DeepFace outputs probabilities across seven standard categories (angry, disgust, fear, happy, sad, surprise, neutral), we report only the dominant predicted label per clip. This clip-level prediction aligns well with human intuition and provides interpretable summaries for downstream

analysis.

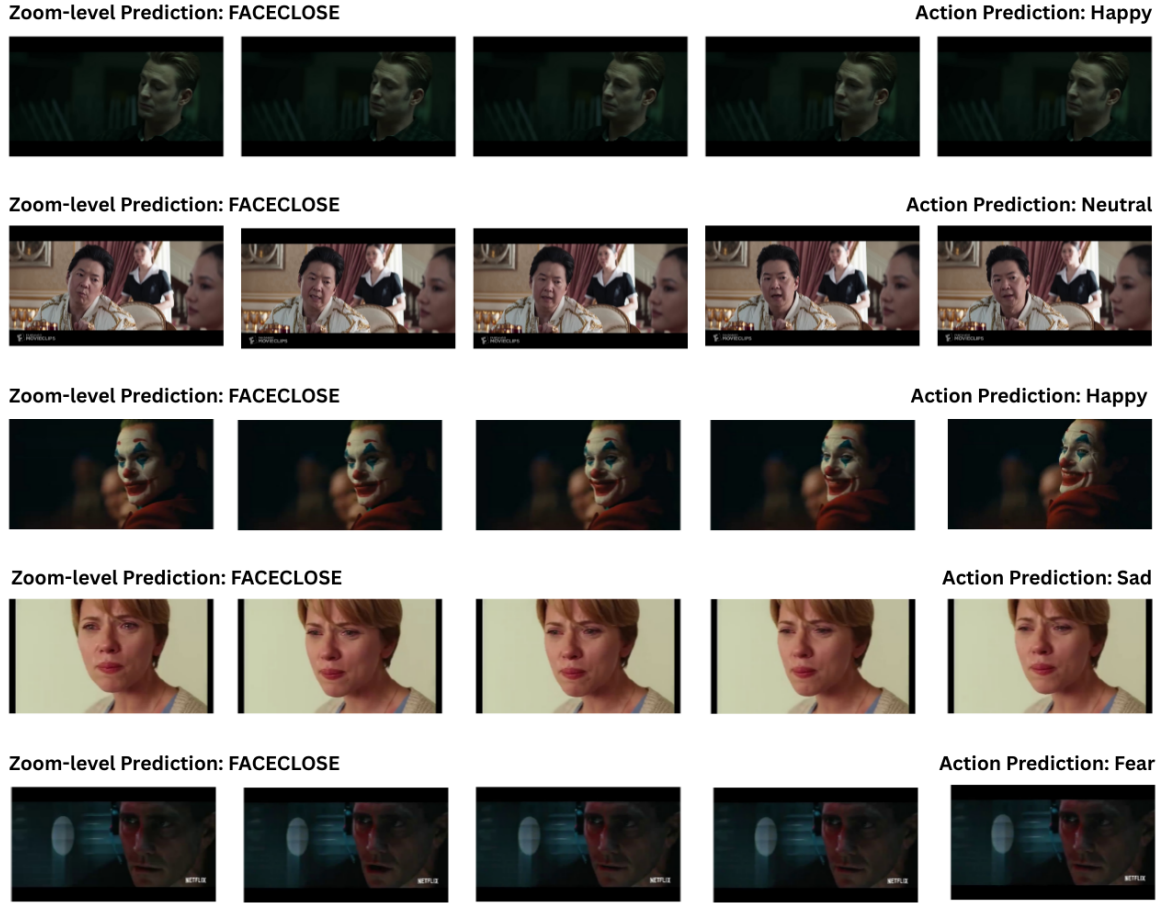


Figure 5.7: Qualitative examples of facial emotion recognition predictions. Each row corresponds to a video clip, where frames are shown alongside the majority-vote predicted emotion label.

5.5.4 Upper-body Action Recognition

Upper-body actions are significant in cinematic material, as they often convey communicative intent through posture, arm gestures, or subtle torso shifts. For this task, we repurposed the [CONFIDANT](#) model, initially introduced in Chapter 4 for near-field gesture

analysis. **CONFIDANT** was explicitly designed to capture **micro-gestures** and confidence-related cues in presenter-style videos, leveraging spatio-temporal attention across cropped upper-body regions. In the present work, we adapt the same architecture to cinematic footage, where medium shots provide comparable torso- and arm-centric information, but under more variable lighting, framing, and stylistic conventions.

For evaluation, we applied **CONFIDANT** on the curated **ZoomGate** dataset, with action segments manually annotated into upper-body categories. The system produced consistent predictions of nuanced upper-body behaviours such as "quick, erratic movements", "turning head to the side" and "arms akimbo". These results underline the discriminative strength of the model, particularly in differentiating subtle postural cues that are frequently overlooked by full-body recognition frameworks. Figure 5.8 presents qualitative predictions obtained from film material.

5.5.5 Full-body Action Recognition

To capture holistic bodily movements and context-dependent gestures, we employed a **STGCN++** model, using pretrained weights on the Kinetics-400 dataset, which encompasses 400 diverse human action classes. This large-scale pretraining provides a substantial prior, allowing the model to generalize effectively to our cinematic dataset without requiring extensive finetuning.

Figure 5.9 presents representative qualitative examples across different action categories. In the first sequence, the model correctly identifies a spraying action in an animated movie, successfully capturing the interaction between the hand, spray can, and wall context. The second sequence demonstrates accurate recognition of squatting, where the model captures lower-body joint dynamics and posture. In the third sequence, however, the model predicts clapping, while the actual scene depicts a boy holding a toy. This illustrates both the model’s sensitivity to repetitive upper-body cues and its occasional misinterpretation of ambiguous poses in cinematic contexts. Overall, these results highlight **STGCN++**’s capacity to parse full-body movements across animated and live-action material, while also pointing to limitations in scenarios with object interactions.

5.6 Conclusion

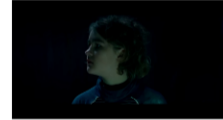
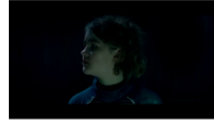
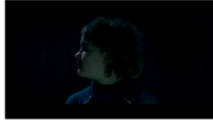
In this chapter, we introduced **ZoomGate**, a unified framework for scale-aware **HAR** in cinematic video. Unlike conventional pipelines that apply a single model to all inputs,

Zoom-level Prediction: UPPERBODY



Action Prediction: Quick, erratic movements

Zoom-level Prediction: UPPERBODY



Action Prediction: Turning head to the side

Zoom-level Prediction: UPPERBODY



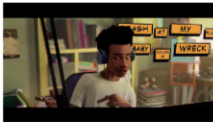
Action Prediction: Turning back to audience

Zoom-level Prediction: UPPERBODY



Action Prediction: Arms akimbo

Zoom-level Prediction: UPPERBODY



Action Prediction: Quick, erratic movements

Figure 5.8: Qualitative results for upper-body action recognition using the [CONFIDANT](#) model. Each row corresponds to a distinct action segment with the predicted action label on the right.



Figure 5.9: Qualitative results of full-body action recognition using [STGCN++](#) pretrained on Kinetics-400. Examples demonstrate correct predictions for spraying and squatting, as well as a misclassification where a boy holding a toy was incorrectly predicted as clapping.

[ZoomGate](#) explicitly disentangles the problem into three stages: (i) assigning zoom-view categories that reflect the visible set of human [keypoints](#), (ii) stabilizing these assignments with temporal smoothing, and (iii) routing coherent segments to specialized recognizers. This design directly addresses the scale variability of film material, where wide shots, medium shots, and close-ups highlight fundamentally different cues.

Our semi-supervised dataset of movie trailers provided a challenging testbed, capturing diverse genres, animated and live-action content, and heterogeneous cinematographic styles. Iterative refinement of the heuristic rules allowed us to identify and resolve systematic failure cases such as tracker instability, occlusions, and far-scale ambiguities. The updated rules, combined with temporal post-processing, yielded stable and interpretable zoom-view labels, which serve both as reliable inputs for downstream recognition and as an interpretable diagnostic layer.

Through qualitative experiments, we demonstrated that specialized models benefit from this scale-aware routing: facial emotion recognition thrives in close-ups, [CONFIDANT](#) effectively captures [micro-gestures](#) in UPPERBODY views, and [STGCN++](#) robustly recognizes whole-body actions in FULLBODY shots. While some misclassifications remain, particu-

larly in ambiguous cases involving [contextual information](#) such as object interactions, the framework consistently improves robustness across mixed cinematic conditions.

Overall, [ZoomGate](#) bridges near-field ([CONFIDANT](#)) and far-field ([HAIKYU](#)) paradigms, showing that scale-awareness is essential for [HAR](#) in real-world film and broadcast material. Beyond performance, the framework contributes an interpretable intermediate representation, aligning computational recognition with cinematographic conventions. We view this as both a practical step toward robust action recognition in heterogeneous media and a conceptual blueprint for future scale-aware recognition systems.

Chapter 6

Conclusion and Future Work

6.1 Summary of Contributions

This thesis has addressed the overarching challenge of robust [HAR](#) across varying camera distances. While most prior work treats [far-field scenario](#), [near-field scenario](#), and [zoom-mixed scenarios](#) in isolation, this research demonstrates that scale-aware preprocessing combined with context-specific skeletons can yield reliable recognition across the full spectrum. The main contributions can be summarized as follows.

First, in the [far-field scenario](#) setting, we introduced [HAIKYU](#), a domain-adapted pipeline for ice hockey broadcast footage, where players appear small, move abruptly, and frequently experience occlusion. Two forms of [contextual information](#) were developed: a temporal [bounding-box normalization](#) strategy that stabilizes player scale and suppresses camera jitter, and a hockey-specific skeleton with [stick keypointss](#). Together, these innovations improved top-1 accuracy from 31% to 64%, highlighting the indispensable role of equipment-aware skeletons in sports action recognition.

Second, in the [near-field scenario](#) setting, we developed [CONFIDANT](#), supported by a novel dataset of 38 [micro-gestures](#) explicitly annotated to distinguish between confident and unconfident behaviours in presenter-style videos. Building on this dataset, we designed a pose-based recognizer capable of detecting subtle confidence-related cues such as folded arms, finger fidgets, or clasped hands. To demonstrate downstream utility, we integrated this recognizer into a diffusion-based generative editing pipeline that transforms unconfident gestures into confident ones, while preserving the speaker’s identity and natural appearance.

Finally, for the mixed-scale domain, we proposed [ZoomGate](#), a lightweight zoom-estimation and routing mechanism designed to handle cinematic and vlog-style footage where shot scale changes dynamically. Based on [BB](#) area and [keypoint](#) density, this mechanism assigns each [tracklet](#) to the specialist model (far-field, near-field, or full-body) best suited to the current scale. Experiments confirm that scale-aware routing, combined with context-specific skeletons, delivers robust recognition performance across heterogeneous zoom conditions.

Taken together, these three case studies establish that scale-aware preprocessing and skeleton augmentation are critical design principles for reliable [HAR](#). Beyond advancing recognition accuracy in challenging domains, the thesis also demonstrates how such recognition can enable downstream applications, from fine-grained sports analytics to generative confidence editing in public speaking. These contributions validate the feasibility of scale-aware action recognition, while also pointing to open challenges that motivate future work.

Statement of intent: In the future, we plan to release code, configuration files, and model checkpoints for the methods introduced here, including the ZoomGate scale estimator, routing/orchestration logic, and the verification GUI ([Appendix D](#)). For datasets, HAIKYU is confidential and will not be released ([Appendix A](#)); CONFIDANT is not planned for public release but may be shared with qualified academic researchers on reasonable request under a data-use agreement that prohibits redistribution and deceptive uses ([Appendix C](#)). All materials will be offered for non-commercial research under an Acceptable Use Policy that restricts profiling, surveillance, and deceptive media generation.

6.2 Future Work

While the presented frameworks make progress toward robust, scale-aware action recognition, several promising avenues for exploration remain open. These can be organized into scenario-specific improvements and broader unified directions.

6.2.1 Far-Field Scenario Improvements

In the [far-field scenario](#) setting, future work should expand the current hockey dataset beyond the six core actions studied in this thesis. Including a more exhaustive taxonomy of tactical and defensive maneuvers, as well as rarer events, would strengthen robustness and applicability. Another important direction is to transition from isolated recognition to

continuous action spotting, where the system can automatically identify the start and end of actions in untrimmed broadcast video. Methodologically, incorporating stereo or multi-camera footage could enable more reliable 3D skeleton estimation, mitigating the challenges of occlusions and camera motion. Finally, extending recognition from individual players to team-level context by integrating puck tracking and modelling inter-player interactions would open the door to higher-level tactical analysis, such as play formation recognition or line-change detection.

6.2.2 Near-Field Scenario Enhancements

For the [near-field scenario](#) scenario, the [CONFIDANT](#) dataset represents a novel contribution, but remains limited in size. Enlarging this dataset with more speakers, diverse demographics, and cross-cultural samples would improve generalization and reduce the subjectivity of confidence labels. Another important direction is multimodal integration, combining visual [micro-gestures](#) with audio prosody, linguistic disfluencies, or even physiological signals, enabling a more holistic approach to confidence detection. Future implementations could also focus on providing real-time feedback, offering speakers live guidance in presentation or virtual reality contexts. Since [micro-gestures](#) vary greatly across individuals, adaptive and personalized models that learn each speaker’s baseline behaviours will be essential for reliable, context-aware feedback.

6.2.3 Mixed-Scale and Zoom-Aware Extensions

In the [zoom-mixed scenarios](#) domain, [ZoomGate](#) has demonstrated the benefits of scale awareness; however, considerable room for improvement remains. The current heuristics, based on [BB](#) area and [keypoint](#) density, while effective, are brittle in complex multi-person scenes or under unusual cinematography. A promising future direction is to replace these heuristics with learned scale-estimation modules, perhaps leveraging [transformers](#) or cross-scale attention mechanisms. Beyond scale classification, architectures that blend global (full-body) and local (face/hand) features adaptively could eliminate the need for explicit routing, maintaining continuity even under ambiguous or rapidly changing zoom conditions. Applying [ZoomGate](#) to unconstrained media such as livestreams, social media, or egocentric video would further test its resilience in real-world, less-structured footage.

6.2.4 Toward Unified Action Recognition

Looking ahead to unified action recognition, the ultimate challenge is to move beyond scenario-specific pipelines toward architectures that seamlessly adapt across scale, viewpoint, and application domain. Advances in contrastive learning and cross-attention mechanisms suggest a path toward models with accurate view and scale invariance, unifying [far-field scenario](#), [near-field scenario](#), and [zoom-mixed scenarios](#) within a single backbone. Such models will also benefit from multimodal fusion, combining skeleton data with [RGB](#) appearance, object trajectories, audio, or even textual transcripts to enrich action semantics, particularly in ambiguous or occluded situations. There is also an exciting opportunity for deeper synergy between recognition and generative editing: recognizers can guide edits, while edits themselves can produce augmented data to retrain and improve recognition models, creating a closed-loop cycle.

In summary, future work should expand available datasets, incorporate richer modalities, replace heuristics with adaptive modules, and strive toward unified frameworks capable of handling the full diversity of human action across scales, viewpoints, and domains. Together, these directions point toward the long-term goal of unified, adaptive, and trustworthy action recognition systems that bridge the gap between analysis and generative intervention.

References

- [1] Image quality metrics: Psnr vs. ssim | ieee conference publication | ieee xplore. Publisher: IEEE.
- [2] Md. Atiqur Rahman Ahad. *Computer Vision and Action Recognition: A Guide for Image Processing and Computer Vision Community for Action Understanding*, volume 5 of *Atlantis Ambient and Pervasive Intelligence*. Atlantis Press, Paris, Dec 2011.
- [3] Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky. Bot-sort: Robust associations multi-pedestrian tracking. *arXiv preprint arXiv:2206.14651*, 2022.
- [4] Bavesh Balaji, Jerrin Bright, Yuhao Chen, Sirisha Rambhatla, John S. Zelek, and David Anthony Clausi. Seeing beyond the crop: using language priors for out-of-bounding box keypoint prediction. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2025. Curran Associates Inc.
- [5] Bavesh Balaji and David A Clausi. Towards agile human pose estimation: A benchmark study in ice hockey analytics. *Journal of Computational Vision and Imaging Systems*, 9(1):54–57, Apr. 2024.
- [6] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, Yanzhen Li, Michael Rubinstein, Tomer Michaeli, Oliver Wang, Deqing Sun, Tali Dekel, and Inbar Mosseri. Lumiere: A Space-Time Diffusion Model for Video Generation, Feb 2024. arXiv:2401.12945.
- [7] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*, July 2021.

- [8] João Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4724–4733, 2017.
- [9] Chia-Chih Chen and J. K. Aggarwal. Recognizing human action from a far field of view. In *2009 Workshop on Motion and Video Computing (WMVC)*, pages 1–7, 2009.
- [10] Haoyu Chen, Henglin Shi, Xin Liu, Xiaobai Li, and Guoying Zhao. Smg: A micro-gesture dataset towards spontaneous body gestures for emotional stress state analysis. *International Journal of Computer Vision*, 131(6):1346–1366, Jun 2023.
- [11] Xinnian Cheng and Zhonghui Li. Research on basketball shooting action recognition and optimization system based on deep learning. pages 546–551, 06 2024.
- [12] Anthony Cioppa, Adrien Deliege, Silvio Giancola, Bernard Ghanem, Marc Van Droogenbroeck, Rikke Gade, and Thomas B Moeslund. A context-aware loss function for action spotting in soccer videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13126–13136, 2020.
- [13] MMAAction2 Contributors. Openmmlab’s next generation video understanding toolbox and benchmark. <https://github.com/open-mmlab/mmaaction2>, 2020.
- [14] PYSKL Contributors. Pyskl: a toolbox for skeleton-based video understanding. <https://github.com/kennymckormick/pyskl>, 2022.
- [15] CUT. Cut youtube channel, 2024.
- [16] Julien Denize, Mykola Liashuha, Jaonary Rabarisoa, Astrid Orcesi, and Romain Hérault. Comedian: Self-supervised learning and knowledge distillation for action spotting using transformers, 09 2023.
- [17] Haodong Duan, Jiaqi Wang, Kai Chen, and Dahua Lin. Pyskl: Towards good practices for skeleton action recognition. In *Proceedings of the 30th ACM International Conference on Multimedia, MM ’22*, page 7351–7354, New York, NY, USA, 2022. Association for Computing Machinery.
- [18] Mehrnaz Fani, Helmut Neher, David Clausi, Alexander Wong, and John Zelek. Hockey action recognition via integrated stacked hourglass network. pages 85–93, 07 2017.
- [19] Harshala Gammulle, David Ahmedt-Aristizabal, Simon Denman, Lachlan Tychsen-Smith, Lars Petersson, and Clinton Fookes. Continuous human action recognition for human-machine interaction: A review. *ACM Comput. Surv.*, 55(13s), July 2023.

- [20] Silvio Giancola, Mohieddine Amine, Tarek Dghaily, and Bernard Ghanem. Soccernet: A scalable dataset for action spotting in soccer videos. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1792–179210, 2018.
- [21] 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.
- [22] Ryo Hachiuma, Fumiaki Sato, and Taiki Sekii. Unified keypoint-based action recognition framework via structured keypoint pooling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22962–22971, 2023.
- [23] Yamin Han, Peng Zhang, Tao Zhuo, Wei Huang, and Yanning Zhang. Going deeper with two-stream convnets for action recognition in video surveillance. *Pattern Recognition Letters*, 107:83–90, 2018. Video Surveillance-oriented Biometrics.
- [24] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS’17)*, pages 6629–6640, Red Hook, NY, USA, 2017. Curran Associates Inc.
- [25] Kristina Host and Marina Ivašić-Kos. An overview of human action recognition in sports based on computer vision. *Heliyon*, 8:e09633, 06 2022.
- [26] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. In *The European Conference on Computer Vision (ECCV)*, 2020.
- [27] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX*, page 196–214, Berlin, Heidelberg, 2020. Springer-Verlag.
- [28] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014.

- [29] Adam Kendon. Gesticulation and speech: Two aspects of the process of utterance. In *The Relationship of Verbal and Nonverbal Communication*, pages 207–228. De Gruyter Mouton, Aug 2011.
- [30] Soroush Khobdeh, Mohammadreza Yamaghani, and Siavash Sareshekeh. Basketball action recognition based on the combination of yolo and a deep fuzzy lstm network, 05 2023.
- [31] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing.
- [32] Xin Liu, Henglin Shi, Haoyu Chen, Zitong Yu, Xiaobai Li, and Guoying Zhao. imigue: An identity-free video dataset for micro-gesture understanding and emotion analysis, Jul 2021. arXiv:2107.00285, CVPR 2021.
- [33] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. Mediapipe: A framework for perceiving and processing reality. In *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019*, 2019.
- [34] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM ’22, page 638–647, New York, NY, USA, 2022. Association for Computing Machinery.
- [35] Mekhail Mustak, Joni Salminen, Matti Mäntymäki, Arafat Rahman, and Yogesh K. Dwivedi. Deepfakes: Deceptions, mitigations, and opportunities. *Journal of Business Research*, 154:113368, Jan 2023.
- [36] OpenAI. ChatGPT: An AI Language Model. <https://chat.openai.com>, 2023.
- [37] Zhaoqing Pan, Weijie Yu, Xiaokai Yi, Asifullah Khan, Feng Yuan, and Yuhui Zheng. Recent progress on generative adversarial networks (gans): A survey. *IEEE Access*, 7:36322–36333, 2019.
- [38] Harish Prakash, Yuhao Chen, Sirisha Rambhatla, David Clausi, and John Zelek. Vip-htd : A public benchmark for multi-player tracking in ice hockey. 02 2024.

- [39] Anik Sen, Syed Hossain, Ashraf Russo, Deb Kaushik, and Kang-Hyun Jo. Fine-grained soccer actions classification using deep neural network. pages 1–6, 07 2022.
- [40] Sefik Serengil and Alper Ozpinar. A benchmark of facial recognition pipelines and co-usability performances of modules. *Journal of Information Technologies*, 17(2):95–107, 2024.
- [41] Mostafa Shahabinejad, Irina Kezele, Seyed Shahabeddin Nabavi, Wentao Liu, Seel Patel, Yuanhao Yu, Yang Wang, and Jin Tang. Video action recognition with adaptive zooming using motion residuals. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 1206–1215, 2023.
- [42] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu. Two-stream adaptive graph convolutional networks for skeleton-based action recognition. In *CVPR*, 2019.
- [43] Moore Threads. Moore-animateanyone. <https://github.com/MooreThreads/Moore-AnimateAnyone>, 2024. GitHub repository.
- [44] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, 2022.
- [45] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 4489–4497, 2015.
- [46] Tuyen Tran, Thao Minh Le, Hung Tran, and Truyen Tran. Unified framework with consistency across modalities for human activity recognition. *arXiv preprint arXiv:2409.02385*, 2024.
- [47] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. FVD: A new metric for video generation, 2019.
- [48] Rejin Varghese and Sambath M. Yolov8: A novel object detection algorithm with enhanced performance and robustness. In *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, pages 1–6, 2024.
- [49] Karan Vats, Hannes Neher, David A. Clausi, and John Zelek. Two-stream action recognition in ice hockey using player pose sequences and optical flows. In *2019 16th Conference on Computer and Robot Vision (CRV)*, pages 181–188. IEEE, May 2019.

- [50] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, Apr 2004.
- [51] Hao Wen, Ziqian Lu, Fengli Shen, Zhe-Ming Lu, and Jialin Cui. Improving skeleton-based action recognition with interactive object information. *International Journal of Multimedia Information Retrieval*, 14(1):3, 2025.
- [52] Feiyi Xu, Feng Xu, Jiucheng Xie, Chi-Man Pun, Huimin Lu, and Hao Gao. Action recognition framework in traffic scene for autonomous driving system. *IEEE Transactions on Intelligent Transportation Systems*, 23(11):22301–22311, 2022.
- [53] Mingwang Xu, Hui Li, Qingkun Su, Hanlin Shang, Liwei Zhang, Ce Liu, Jingdong Wang, Yao Yao, and Siyu zhu. Hallo: Hierarchical audio-driven visual synthesis for portrait image animation, 2024.
- [54] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. Vitpose: Simple vision transformer baselines for human pose estimation, Oct 2022. arXiv:2204.12484.
- [55] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition, Jan 2018. arXiv:1801.07455, version: 2.
- [56] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation, Aug 2023. arXiv:2307.15880.
- [57] Jaehong Yoon, Shoubin Yu, and Mohit Bansal. RACCooN: A Versatile Instructional Video Editing Framework with Auto-Generated Narratives, Oct 2024. arXiv:2405.18406.
- [58] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, Jun 2018.
- [59] Yuang Zhang, Jiayi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance, Jun 2024. arXiv:2406.19680.
- [60] He Zhu, Junwei Liang, Chengzhi Lin, Jun Zhang, and Jianming Hu. A transformer-based system for action spotting in soccer videos. pages 103–109, 10 2022.

APPENDICES

Appendix A

Dataset Card: HAIKYU

Summary. 1,547 two-second NHL broadcast clips at 30 fps centered on six actions, with person tracklets, per-frame bounding boxes, and 2-D keypoints from a hockey-tuned pose model; temporal box normalization as in Chapter 3.

Composition and splits. 29 teams, male adult professional athletes; stratified 70/30 train/validation split with per-class reporting due to imbalance.

Intended use. Research on far-field sports action recognition and keypoint modeling; not representative of amateurs or youth leagues.

Out-of-scope use. Surveillance, identity profiling, or commercial broadcast replacement.

Collection and preprocessing. Public broadcast sources; tracklets via VIP-HTD [38]; hockey-specific keypoints; temporal normalization.

Demographics and consent. Professional male adult athletes; to our knowledge no minors.

Licensing and availability. *Confidential:* due to rightsholder/licensing constraints, HAIKYU will not be publicly released; neither raw videos nor derived annotations will be distributed. This thesis provides aggregate statistics and methodological details for independent reproduction on lawfully obtained broadcasts.

Known limitations. Broadcast motion blur, class imbalance, and narrow domain (rink/broadcast angles).

Appendix B

ChatGPT Classification for CONFIDANT Dataset

The classifications provided by ChatGPT are summarized in the Table [B.1](#). From the classifications, it is clear that gestures that involve open body language, upright posture, and controlled movements, such as standing tall, hands behind the back, arms akimbo, starting arm movements from the shoulders, open hands, smooth movements, head up, and the palm-to-palm (Namaste) gesture, are associated with confidence. In contrast, gestures involving self-touching, fidgeting, closed body language, or avoidance behaviors, such as crossing fingers, folding arms, fidgeting fingers, covering the mouth, playing with clothes, head down, touching or scratching facial parts, and putting hands in pockets, are linked to a lack of confidence.

Table B.1: Classification of Micro-Gestures Using ChatGPT

Micro-Gesture	Classification
Crossing fingers	Unconfident
Standing tall with an upright posture	Confident
Hands behind the back	Confident
Quick, erratic movements	Unconfident
Folding arms	Unconfident
Fidgeting fingers	Unconfident
Arms akimbo	Confident
Covering mouth	Unconfident
Playing with clothes	Unconfident
Head down	Unconfident
Showing the back of the palms	Unconfident
Holding objects	Unconfident
Turning back to audience	Unconfident
Covering chest	Unconfident
Playing or adjusting hair	Unconfident
Rubbing or holding hands	Unconfident
Turning head to the side	Unconfident
Touching or scratching facial parts	Unconfident
Moving torso	Unconfident
Starting arm movements from shoulders	Confident
Keeping hands open and visible	Confident
Touching jaw	Unconfident
Putting hands in pockets	Unconfident
Touching or scratching head	Unconfident
Playing or manipulating objects	Unconfident
Finger-pointing gestures	Unconfident
Shaking shoulders	Unconfident
Smooth and controlled movements	Confident
Deep breath	Unconfident
Head up	Confident
Scratching or touching arms	Unconfident
Covering face	Unconfident
Touching or scratching neck	Unconfident
Rubbing eyes	Unconfident
Touching nose	Unconfident
Palm-to-palm (Namaste gesture)	Confident
Touching or scratching forehead	Unconfident
Touching hat	Unconfident

Appendix C

Dataset Card: CONFIDANT

Summary. 412 near-field clips from presenter-style online videos, annotated into 38 upper-body micro-gesture classes and mapped to {Confident, Unconfident}; upper-body keypoints provided.

Composition and splits. Speakers appear to be consenting adults of mixed genders/ethnicities (not systematically recorded); stratified 70/30 train/validation split.

Intended use. Research on micro-gesture recognition and *consented* behavior editing.

Out-of-scope use. Deceptive or manipulative edits; hiring/credit screening; biometric identification; surveillance.

Collection and preprocessing. Presenter segments curated from public sources; DWPose [56] with upper-body joints retained; two skeleton configurations evaluated.

Licensing and availability. No public release planned at this time; however, the dataset may be shared on reasonable request with academic researchers under a data-use agreement prohibiting redistribution and deceptive uses. Any shareable package will include annotations (timestamps, labels, and video source). Requests: *kbuzko@uwaterloo.ca*.

Known limitations. Small size, partial cultural coverage of 'confidence', and overlap between gesture classes.

Appendix D

Resource Note: ZoomGate

Summary. ZoomGate is a scale-aware action recognition pipeline that detects frame-wise shot scale and switches to a scale-appropriate action recognition expert.

Sourcing and pseudo-labeling. We collected candidate clips from YouTube by querying for trailers with frequent zooms. We generated pseudo-labels by running ScaleGate, extracting person tracklets, and routing them to specialist recognizers. A light manual audit of random samples was used to estimate precision and prune obvious errors.

Inputs and preprocessing. Person tracklets with per-frame boxes; optional near-field crops; 2-D keypoints.

Outputs. Per-tracklet action labels with associated zoom scale.

Intended use. Research on zoom-mixed (mixed-zoom) action recognition and scale-aware routing.

Out-of-scope use. Surveillance, identity profiling, and deceptive media generation.

Licensing and availability. We intend to release code, configuration files, and the verification GUI for non-commercial research.

Known limitations. Pseudo-labels can contain noise and inherit domain bias from source videos; errors in detection/tracking and scale estimation can propagate; trailer and presenter domains may not generalize to handheld/social video without adaptation.

Contact. Code/usage questions: kbuzko@uwaterloo.ca.

Appendix E

Pseudocode Listings

This appendix contains pseudocode listings for algorithms described in Chapter [5](#). We include complete rule-based view classification logic (ZoomGate), allowing for independent reproduction of the implementation, regardless of the source code.

Algorithm 1: ZoomGate Updated View Classification

Input: face_bb, body_bb, lhand_bb, rhand_bb, torso_vis_count, lower_seen, frame_area, frame_h, W, H
Output: Zoom-view label

face_ok $\leftarrow$ (face_bb $\neq \emptyset$);
face_area_frac $\leftarrow$ Area(face_bb) / frame_area;
face_h_frac $\leftarrow$ Height(face_bb) / frame_h;
hands_ok $\leftarrow$ ValidHand(lhand_bb, body_bb) $\vee$ ValidHand(rhand_bb, body_bb);
pose_ok $\leftarrow$ (torso_vis_count > 0) $\vee$ lower_seen;
if not (face_ok $\vee$ pose_ok $\vee$ hands_ok) **then**
 if EdgeTouch(body_bb, W, H) ≥ 2 **or** BlurGuard(body_bb) **then**
 return NOPARTS;
if face_ok **and** (face_area_frac > 0.07 $\vee$ face_h_frac > 0.30) **then**
 if torso_vis_count < 2 **then**
 return FACECLOSE
 else
 prefer UPPERBODY
if hands_ok **and not** (face_ok $\vee$ torso_vis_count ≥ 2 $\vee$ lower_seen) **then**
 return HANDSONLY;
if body_bb $\neq \emptyset$ **then**
 body_h_frac $\leftarrow$ Height(body_bb)/frame_h;
 if lower_seen **and** (face_ok $\vee$ torso_vis_count ≥ 2) **then**
 return FULLBODY
 else if body_h_frac < 0.30 **and** (not face_ok $\vee$ face_h_frac < 0.12) **then**
 return FULLBODY // far-person fallback
 else if face_ok $\vee$ torso_vis_count ≥ 2 **then**
 return UPPERBODY
 else
 return NOPARTS
if face_ok **then**
 return UPPERBODY
if hands_ok **then**
 return HANDSONLY
return NOPARTS;

Appendix F

Hand Gesture Label Set

Table F.1 lists the final set of hand gesture labels adapted for cinematic scenarios.

Communicative	Expressive / Emotional	Contextual / Symbolic
Wave	Fist (anger, power)	Make a phone call
Beckon	Hold fist in the other hand	Framing gesture (photo frame)
Stop / palm-out	Push away	Prayer / pleading
Thumbs up	Snap fingers	Heart (dual hands / fingers)
Thumbs down	Applaud	Counting (showing numbers)
Like / dislike	Palm-to-fist (determination)	OK sign
Peace / victory	Pointing / index finger	Rock sign

Table F.1: Adapted hand gesture labels for cinematic scenarios.

Glossary

bounding-box normalization Resizing detected person bounding boxes in a tracklet to a consistent scale to reduce pose jitter caused by camera motion or zoom [iv](#), [5](#), [6](#), [18–20](#), [22](#), [23](#), [25–28](#), [63](#)

COCO Common Objects in Context dataset; includes a 17-keypoint human skeleton standard [8](#), [9](#), [19](#), [22](#), [23](#), [25](#), [28](#), [33](#)

COCO-WholeBody An extension of COCO with 133 keypoints (17 body, 68 face, 42 hands, 6 feet) [9](#), [33](#), [42](#)

confidence score A scalar in $[0, 1]$ output by the pose estimator indicating reliability of a detected [keypoint](#) [33](#)

contextual information Scenario-specific cues beyond joint coordinates (e.g., camera scale, equipment, object interactions) that disambiguate actions [1](#), [10](#), [14](#), [17](#), [18](#), [20](#), [28](#), [50](#), [62](#), [63](#)

DWPose A whole-body pose estimation approach designed for accurate, detailed keypoints [33](#), [38](#)

far-field scenario Setting where the subject appears small and distant, emphasizing whole-body motion and contextual cues [iv](#), [viii](#), [ix](#), [1–8](#), [10](#), [11](#), [14](#), [17–19](#), [28](#), [44](#), [45](#), [47](#), [63](#), [64](#), [66](#)

human pose estimation Detecting key anatomical landmarks (joints/limbs) on the human body in images or videos [8](#), [33](#), [43](#)

keypoint A detected anatomical landmark (e.g., joint or facial feature) used in pose estimation [iv](#), [v](#), [1](#), [4–9](#), [12](#), [13](#), [19](#), [21–24](#), [26](#), [28](#), [33](#), [34](#), [37–39](#), [42](#), [46](#), [48](#), [50](#), [51](#), [53](#), [61](#), [64](#), [65](#), [82](#)

micro-gestures Subtle, brief movements (e.g., folding arms, crossing fingers) indicative of psychological states [iv](#), [2–7](#), [12](#), [13](#), [16](#), [17](#), [29](#), [30](#), [33](#), [34](#), [37](#), [38](#), [42–44](#), [48](#), [49](#), [59](#), [61](#), [63](#), [65](#)

near-field scenario Setting where the subject is close to the camera, emphasizing fine-grained gestures and facial expressions [iv](#), [viii](#), [ix](#), [1–8](#), [10](#), [12](#), [14](#), [17](#), [29](#), [33](#), [42–45](#), [47](#), [63](#), [65](#), [66](#)

optical flow A technique to estimate pixel-wise motion between consecutive video frames, often used as a cue for action recognition [10](#)

point cloud A set of points in 3D (or spatiotemporal) space used to represent geometry or trajectories [13](#)

scale gate A lightweight module that computes a zoom/scale score (e.g., from bounding-box area and keypoint density) to route inputs to the appropriate branch [14](#)

scale-aware routing Dispatching a detected person/tracklet to a specialist recognizer based on estimated camera scale or keypoint density [29](#), [44](#)

skeleton-based action recognition Recognizing actions from sequences of body keypoints, abstracting appearance and background [8](#), [9](#), [14](#), [32](#), [42](#)

stick keypoints Sport-specific landmarks (e.g., butt end, heel, toe) added to the skeleton to capture equipment pose and motion [17–19](#), [21–23](#), [25](#), [28](#), [63](#)

tracklet A short, temporally continuous sequence of detections corresponding to the same subject, typically produced by a tracker [v](#), [5–7](#), [14](#), [17](#), [19–22](#), [45](#), [48](#), [53](#), [64](#)

transformer A neural architecture using self-attention to model long-range dependencies in sequences [56](#), [65](#)

zoom-mixed scenario Setting where the camera zoom or editing abruptly shifts shot scale, requiring models to adapt between far-field and near-field cues within a single video [iv](#), [v](#), [ix](#), [2](#), [3](#), [5–8](#), [13–15](#), [17](#), [45](#), [49](#), [63](#), [65](#), [66](#)