國立中興大學資訊工程學系

碩士學位論文

結合骨架時序轉換器與球拍幾何特徵之雙模態深度神經網路於桌球擊球動作辨識之研究

A Novel Table Tennis Stroke Recognition Method Using The Bimodal Deep Neural Networks with Skeletal-Temporal Transformer and Racket Geometric Features

指導教授：吳俊霖　　Jiunn-Lin Wu

研 究 生：黃照恩　　Chao-En Huang

中華民國　一一四　年　八　月

國立中興大學 資訊工程學系

碩士學位論文

題目：結合骨架時序轉換器與球拍幾何特徵之雙模態深度神經網路於桌球擊球動作辨識之研究

A Novel Table Tennis Stroke Recognition Method Using The Bimodal Deep Neural Networks with Skeletal-Temporal Transformer and Racket Geometric Features

姓名：黃照恩　學號：7112056111

經　口　試　通　過　特　此　證　明

論文指導教授 吳俊霖

論文考試委員 林惠勇
許銘勇
吳俊霖

中華民國　114　年　7　月　14　日

致謝

離開職場後重返校園，是我人生中一段很大膽的選擇。這兩年的研究所生活充滿挑戰，也帶來許多轉變與成長。如今走到畢業的這一刻，我懷著無比感激的心情，想謝謝那些在這段旅程中支持、陪伴與鼓勵我的人，獻上我最誠摯的感謝。

首先，衷心感謝我的指導教授吳俊霖老師。在研究方向的確立、實驗過程的討論，以及論文撰寫的各個階段，老師都給予我悉心指導。其次，也想要感謝我的父母，謝謝你們始終支持我做出的每一個選擇，給予我理解、鼓勵，並在經濟上提供幫助，讓我能安心完成這份學業。

此外，我要特別感謝智絜。在這段研究所的旅程中，謝謝妳始終如一地陪伴著我，讓我的生活因妳而豐富又溫暖。還記得我們經常聊天到日出，到望高寮談論彼此的想法與心情。曾一起走進忘憂谷，把煩悶和苦惱拋諸腦後，也在日常裡煮飯、做甜點，讓平凡的每一天都像節日一樣。我們走遍台北、台中、彰化、台南的大街小巷，品嚐過無數美食，也一同度過許多節日與特別的時刻。從新手駕駛的興奮與緊張，到經歷車禍的驚慌與互相扶持，每一段經歷都深深烙印在我的心裡。也謝謝妳在我生病時的細心照顧，在我低潮時給予理解與鼓勵，陪我走過每一段難熬的時光。從一起挑禮物、寫卡片，到用相機記下那些小小的回憶，再到陪我走到畢業的這一天，我們一起經歷的每個片段，都讓這段求學歲月變得特別而難忘。

謝謝妳，為我平凡的研究所生活添上了色彩，讓每一天都閃耀著光芒。這段求學之路，不再只是追求知識的過程，也因為有妳，多了陪伴、理解與溫柔的力量。未來的路上，我期盼仍有妳的陪伴，繼續肩並肩走向人生的下一段旅程。

最後，向所有曾經幫助和支持過我的人致以最誠摯的感謝，你們的每一份陪伴與支持，我都深深銘記在心。

摘要

隨著電腦視覺技術的快速發展，動作辨識已成為智慧體育分析中的關鍵應用之一，尤其在技術精細且動作快速的運動項目如桌球中，更需仰賴模型對細節姿態與時間脈絡的高度理解能力。為因應此挑戰，本研究使用一套雙模態動作辨識架構，整合具備時間建模能力之 2D 骨架資訊與具豐富外觀特徵的 RGB 視覺資訊，強化模型在桌球擊球任務中的分類效能與時序辨識能力。骨架模態部分，採用 SkateFormer 模型以捕捉跨時間的骨架變化與語意關聯，克服傳統 Holistic Interaction Transformer 僅採用單幀姿態所面臨的時序侷限與方向性模糊問題。RGB 模態則引入 SlowFast ResNet 結構，擷取動作中的紋理細節與背景脈絡，以補足骨架模態於關節遮蔽或估測錯誤情境下的資訊不足。此外，本研究引入球拍的區域面積與中心座標作為幾何特徵，以輔助模型掌握揮拍動作的起始與結束時機，改善邊界階段的分類錯誤。實驗使用由專家示範所建構之高解析度桌球擊球資料集，涵蓋正手與反手共八類典型擊球動作，包括反手切球、反手擰球、反手推球、反手拉球，以及正手切球、正手平擊、正手殺球、正手拉球，並以滑動視窗方式生成具時間連續性的標註片段進行訓練與評估。結果顯示，本方法於 Precision、Recall 與 F1-score 上分別達到 96.1%、96.4% 與 96.2%，相較 HIT Network 相對提升了 25.9%，整體效能明顯優於 HIT Network、SkateFormer 及 SlowFast ResNet 等基準模型。同時在 JHMDB 通用動作資料集上亦取得 84.2%、83.8%、83.8% 的分類表現，相較 HIT Network 相對提升了 1.7%，展現良好的泛化能力。本研究證實骨架與影像模態的互補性，及幾何特徵對動作辨識的幫助，展現其在實務應用上的潛力。

關鍵字：雙模態深度神經網路、桌球擊球動作辨識、骨架時序轉換器、球拍幾何特徵、球拍實例分割

Abstract

We used a dual-modality framework for fine-grained action recognition in table tennis, combining skeletal motion and RGB appearance cues to improve both classification accuracy and temporal consistency. The skeletal branch utilizes SkateFormer to model joint dynamics over time, effectively addressing the temporal limitations and directional ambiguity observed in prior methods such as the Holistic Interaction Transformer, which relies on single-frame poses. Complementing this, the RGB branch employs SlowFast ResNet to extract texture and context features, providing resilience against occlusions and pose estimation errors.

To further enhance boundary recognition, where misclassifications often occur, we incorporate geometric cues from racket segmentation, such as area and center coordinates, enabling the model to better infer stroke onset and offset phases.

Experiments on a high-resolution expert-annotated table tennis dataset covering eight stroke types demonstrate that our method achieves 96.1% precision, 96.4% recall, and 96.2% F1-score, representing a 25.9% relative improvement over the HIT Network and significantly outperforming state-of-the-art baselines, including HIT Network, SkateFormer, and SlowFast ResNet. Additional evaluation on the general-purpose JHMDB dataset yields 84.2%, 83.8%, and 83.8% respectively, showing a 1.7% relative improvement over the HIT Network and indicating strong generalization capability.

Our results highlight the value of integrating complementary modalities and leveraging task-specific geometric features to advance fine-grained action recognition in real-world sports applications.

Keywords — *Bimodal Deep Neural Network, Table Tennis Stroke Recognition, Skeletal-Temporal Transformer, Racket Geometric Features, Racket Instance Segmentation*

目錄

圖目錄

表目錄

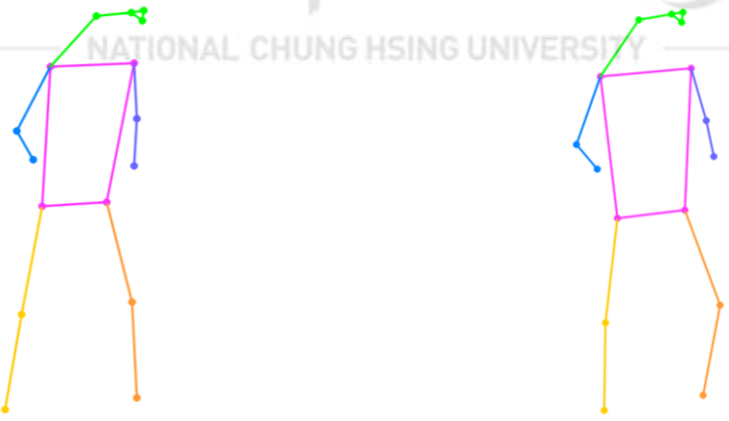
第一章 緒論

1.1 研究背景與動機

近年來，智慧科技與運動分析的結合逐漸興起，加速競技運動向科技化與精準化轉型。在實務層面，傳統桌球訓練或賽後分析多仰賴人工標記與影片回放，不僅繁瑣耗時，亦高度依賴教練經驗，難以量化動作特徵或系統比較擊球策略。特別在處理大量影片進行訓練紀錄與動作統計時，此方式更顯效率低落，且不易捕捉細微技術差異。

為提升動作辨識的自動化與客觀性，近年來研究者開始導入基於深度學習的人體姿態估測 (Human Pose Estimation) [1] 技術，透過卷積神經網路預測關節的二維位置，構建骨架作為輸入以進行動作分類。此類骨架表徵具結構清晰與抗背景干擾等特性，已廣泛應用於運動動作分析任務中。

然而，僅依賴骨架模態進行辨識仍存在多項挑戰。桌球擊球動作間常呈現外觀相似、差異細微的特性，僅從單幀骨架進行判斷，容易忽略關節協調與動作時序等語意線索，導致分類混淆，如圖 1 所示。



(a) (b)

圖 1 外觀相似的單幀骨架姿態對應不同擊球類型

(a) 正手切球 (b) 正手平擊

再者，在實際拍攝環境中，選手動作快速且常伴隨身軀遮擋，使部分關鍵關節因被自身身體遮住而無法準確預測，進而導致骨架點缺失或錯置，影響動作辨識的穩定性與準確度，如圖 2 所示。

(a) (b)

圖 2 骨架點因遮擋產生缺失現象

(a)原始 RGB 影像 (b)人體姿態估測結果

為克服上述限制，近年研究逐漸關注多模態訊息的整合。在模態選擇上，RGB 影像提供了豐富的場景與外觀資訊，能夠補足骨架在遮蔽條件下的關鍵特徵。然而，其也易受到背景雜訊與視角變化影響，辨識表現不易穩定。相較之下，人體骨架能以簡潔的關節表徵去除冗餘背景干擾，強調人體運動的結構特性，是捕捉動作語意的理想資料來源。但若忽略時間層次與動作演化過程，僅依賴靜態姿態輸入，仍難以區辨外觀相似但運動軌跡與發力方式截然不同的動作類型。

此外，擊球語意的判斷不僅仰賴骨架關節的動態軌跡，亦與球拍於擊球瞬間的幾何位置與運動方向密切相關。特別是在動作起始與終止的時間點上，球拍的空間位置與面積往往能提供關鍵線索，輔助判斷擊球是否發生。如圖 3 所示，不同的揮拍方向與角度不僅會改變球體的旋轉形式與飛行軌跡，更直接體現了選手所採用的擊球姿勢與技術策略。透過追蹤球拍區域的幾何特徵，可望提升對動作起迄點的辨識準確性，進一步強化整體擊球語意的時序建模能力。

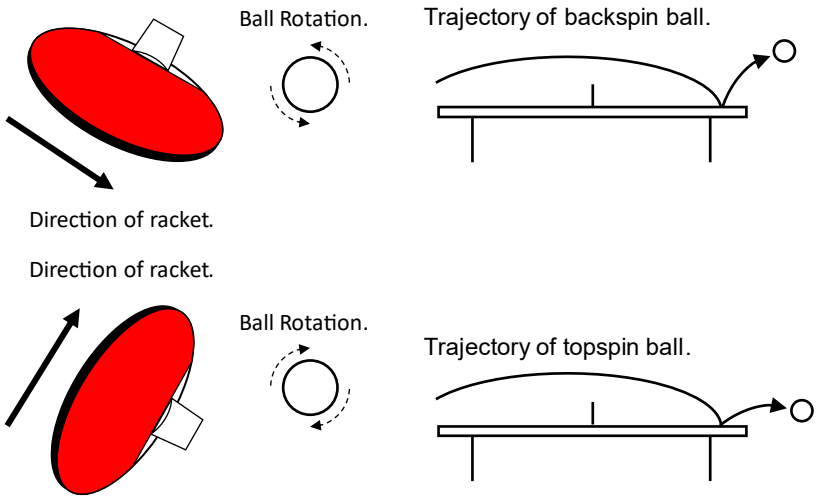


圖 3 不同揮拍方向與角度反映擊球策略與球路變化

基於此觀察，本研究進一步整合骨架模態中的空間與時間資訊，採用具時序建模能力之架構以捕捉動作連貫性，同時引入 RGB 模態作為輔助，以補強骨架在遮蔽與細節辨識上的侷限，構成雙模態融合設計。此外，本研究亦納入球拍區域的幾何特徵 (如面積與中心位置)，藉此強化對擊球動作起始與結束時間點的掌握。透過骨架與 RGB 的語意互補，以及球拍訊息在時間定位上的輔助，可望建構一套兼具辨識準確性與實務應用潛力的智慧桌球動作辨識系統。

1.2 論文架構

本篇論文的主要架構分成五個章節，各個章節的架構與內容分別如下所述。第一章為緒論，說明本研究的背景與動機，並簡要介紹論文內容安排。第二章為文獻回顧，針對桌球擊球辨識與一般性人類動作辨識之相關研究進行整理與分析，歸納各方法之優劣與應用侷限。第三章為研究方法，詳細說明本研究之動作辨識系統設計，包括球拍影像分割、骨架估計、時序建模與雙模態融合等模組。第四章為實驗結果與討論，呈現本研究在資料集與評估指標下的效能表現，並與既有方法進行比較分析。第五章為結論與未來發展，總結本研究成果，並為後續可延伸的研究方向與應用建議。

第二章 文獻回顧

雖然桌球是一項廣受關注的競技運動，針對其影片中擊球動作之識別研究仍相對有限。為系統性回顧相關方法，本章將文獻依據應用特性分類為桌球擊球辨識之專用方法與人類動作辨識之通用方法兩類，並分別進行探討與比較。

2.1 桌球擊球辨識之專用方法 (Domain-Specific Approaches for Table Tennis Stroke Recognition)

當前針對桌球擊球辨識的研究，主要關注於如何精準擷取動作關鍵特徵。部分方法聚焦於人體姿態的時間變化，亦有方法結合多種視覺模態以提升識別準確度。

2.1.1 基於二維姿態估計的方法 (Approach Based on 2D Pose Estimation)

二維姿態估計 (2D Pose Estimation) [1] 技術廣泛應用於動作分析領域，其核心在於透過電腦視覺模型自影像中擷取人體各部位之關節點 (如頭部、手肘、膝蓋等)，以構成表徵人體骨架的幾何結構資訊。透過這些關節點在時間序列上的變化，可以進一步進行動作分類與識別，尤其在運動分析中已展現出實用價值。

Kulkarni 和 Shenoy 便提出一種結合二維姿態估計與時間卷積網路 (Temporal Convolutional Network, TCN) [2] 動作識別系統，應用於桌球擊球姿勢的分類任務，其架構如圖 4 所示。

他們首先利用 Single Shot Multibox Detector (SSD) [3] 偵測影片中的選手，接著透過 High-Resolution Net (HRNet) [4] 擷取人體 2D 骨架，聚焦於右手肘、右手腕與雙肩四個關鍵部位的連續座標序列作為模型輸入，並以 TCN 模型捕捉這些關節在 100 幀中的時間關聯，以分類十一種常見擊球動作，平均辨識準確率達 98.72%，展示了 2D 骨架資訊在桌球動作識別上的可行性。

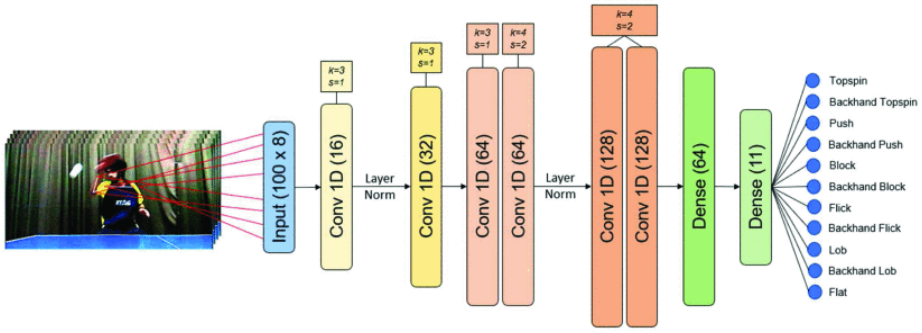


圖 4 TCN 之模型架構圖[2]

然而，該方法仍存在若干限制。首先，其輸入特徵僅涵蓋少數特定關節，未能完整利用人體骨架的整體結構資訊，難以建構全身關節之間的高階語意關聯。其次，TCN 雖可捕捉長距依賴，但其為逐層堆疊式的時序卷積架構，無法靈活選擇動作過程中最具辨識性的關節與關鍵幀 (key joints and key frames)，因此對於高相似度動作的區辨能力有限。此外，其設計未針對空間 (骨架拓撲) 與時間 (動作演化) 進行結構化建模，僅將關節點視為序列資料進行處理，亦可能導致重要語意在時間壓縮過程中遺失。

2.1.2 基於雙分支時空卷積的方法 (Approach Based on Twin Spatio-Temporal Convolutional Networks)

在複雜的體育動作辨識任務中，單一模態往往難以全面捕捉動作的空間與時間特徵。為此，雙模態導向的方法強調結合來自不同資料來源的表徵資訊，藉由互補模態的協同運作，提供更具魯棒性與語意豐富性的辨識能力。這種策略特別適用於桌球這類對細部肢體變化與時間對齊要求極高的場景。

Martin 等人 [5] 所提出的 Twin Spatio-Temporal Convolutional Neural Network (TSTCNN)，即為雙模態導向方法的代表性架構。該模型包含兩條對稱的 3D 卷積分支，分別處理 RGB 原始畫面與其對應之光流 (Optical Flow) 特徵。透過三維卷積操作，TSTCNN 能夠從各模態中同時擷取動作的空間構型與時間動態，最後再將兩者特徵進行融合，以作為分類依據，其架構如圖 5 所示。

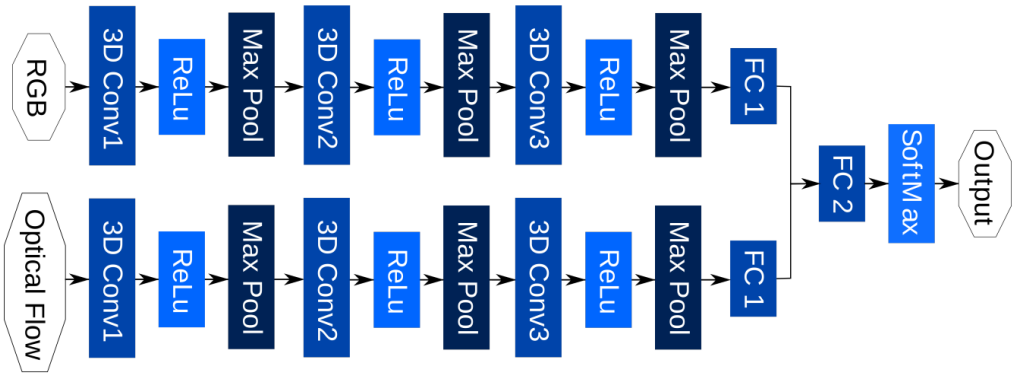
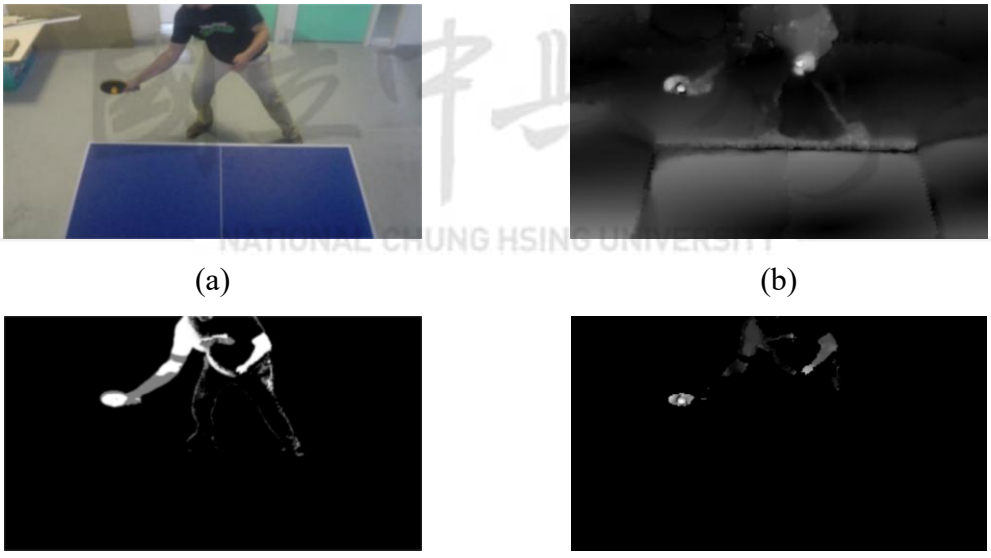


圖 5 TSTCNN 之模型架構圖[5]

三維卷積網路可直接作用於時間軸上的連續畫面，有效捕捉動作的變化趨勢與時序關聯；而光流模態則著重於推估像素間的運動向量，強化模型對動作速度與方向變化的感知。光流法假設相鄰畫面間的像素亮度變化微小，藉此比對像素值以計算其運動，廣泛應用於運動分析、目標追蹤與動作偵測等場景。Martin 等人即採用 Zivkovic 和 Van der Heijden [6] 所提出的方法來提取光流資訊，如圖 6 所示。此種雙模態併用策略，能在含有顯著運動變化的片段中提供更豐富的特徵表示，進一步提升模型對動作區段與類型的辨識準確度。



(a) (b)

(c) (d)

圖 6 光流影像產生流程示意圖[5]

(a)原始 RGB 影像 (b)光流幅度影像

(c)估計出的前景影像 (d)過濾後的光流影像

然而，TSTCNN 也存在數項挑戰與限制。首先，光流的品質對於模型辨識效果具有高度影響，但在桌球這類高速度、強運動模糊的場景下，手部與球體的運動軌跡易出現訊號損失，使得光流模態對快速動作的支持效果有限。

再者，該方法未引入人體結構資訊 (如骨架)，導致模型對於肢體語意與跨關節變化的理解力不足，難以處理僅在肢體角度與關節速度上有所差異的細粒度擊球姿勢。此外，兩分支網路在整個流程中獨立運作，僅於最終階段進行靜態特徵融合，缺乏跨模態間的動態對齊與語意互補機制，可能導致關鍵時序訊息的對齊錯位或語意模糊，限制了其在細緻動作辨識上的表現能力。

2.2 人類動作辨識之通用方法 (General-Purpose Methods for Human Action Recognition)

在人類動作辨識領域中，近年研究多聚焦於骨架隨時間的變化建模，以及多模態資訊間的互動融合，這兩種方向在提升辨識效能上展現出顯著潛力。

2.2.1 基於骨架時序建模的方法 (Approach Based on Temporal Skeleton Modeling)

骨架模態因其天然去除背景與光照等干擾，在動作辨識領域中逐漸成為重要研究方向。Do 與 Kim 所提出的 SkateFormer [7] 為其中代表性成果，其設計以 Transformer 架構 [8] 為基礎，針對骨架資料中不同的空間與時間語意結構 (如鄰近關節與遠距關節、局部時間與全域時間) 進行四分支的劃分與建模，並利用 Partition-Reversal 機制將特徵還原回原始形狀以利融合。此設計不僅提升了長距離依賴建模的能力，也針對骨架時序動作中細節變化具備良好的辨識效果，特別是在動作類型差異明顯、骨架資訊完整的情況下展現優勢。

如圖 7 所示，SkateFormer 的輸入骨架資料是每一幀皆包含完整且準確的關節點資訊。然而，在實際應用中，關節點偵測可能不完整。

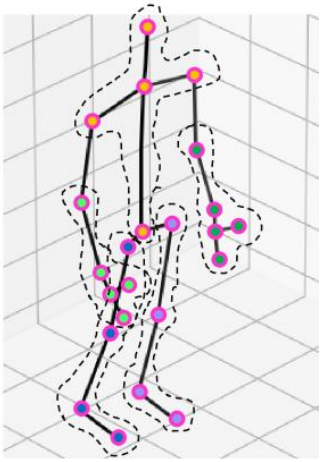


圖 7 理想情況下的完整關節量測[7]

在高動態運動場景中，如桌球選手在進行快速揮拍或轉身等動作時，肢體極易因運動模糊或自身遮擋 (self-occlusion) 而導致骨架估計品質下降。問題包含關節之缺失 (missing joints)、錯置 (misplacement)，或在時序上產生抖動與不連續現象。

這些骨架品質問題在 SkateFormer 架構下將被直接傳遞至注意力機制中，影響模型對語意結構的聚焦能力。對於需要細緻區辨動作變化的應用場景 (如分辨類似擊球形式)，這些骨架輸入誤差更會使模型難以聚焦於實際發生語意變化的關節部位，導致辨識結果不穩定。更關鍵的是，SkateFormer 完全依賴骨架模態進行建模，並未引入其他模態 (如 RGB 或光流) 進行輔助對齊與語意補強，當骨架品質不穩或關節點遺失時，便缺乏額外線索來矯正或補充錯誤，進一步放大模型對低品質骨架的敏感性。

2.2.2 基於互動注意力建模的方法 (Approach Based on Interaction-Aware Attention Modeling)

在動作辨識任務中，建構跨時間與空間的互動關係，對於正確理解複雜動作語意至關重要。Faure 等人提出的 Holistic Interaction Transformer Network (HIT Network) [9] 即是一種以互動為核心的注意力建模架構，透過精心設計的雙分支 Transformer 模組，同時捕捉 RGB 與骨架模態下的人體動作特徵。

RGB 分支擅長捕捉外觀與動態資訊，對於整體動作過程中的視覺變化與物件互動具有良好辨識能力；而骨架分支則專注於建構基於人體關節點的空間關聯，補足外觀特徵難以解析的動作語意。兩者最終透過 Attentive Feature Fusion Module 進行深層特徵融合，使得模型得以整合不同模態的優勢，在動作辨識與時間區段預測 (temporal action localization) 上展現良好效果。

然而，如圖 8 所示，HIT Network 在骨架模態的處理上僅針對單幀進行 Pose Encoding，在骨架模態上缺乏時間維度的資訊擷取與建模。此限制導致模型在辨識動作過程中高度相似但語意細節不同的類別時容易混淆，特別是需要依賴動作演變順序 (如肢體運動方向或速度變化) 進行區辨的情況。由於無法捕捉跨幀的骨架變化與時序依賴，進而限制其於高動態或細節敏感動作上的辨識表現。

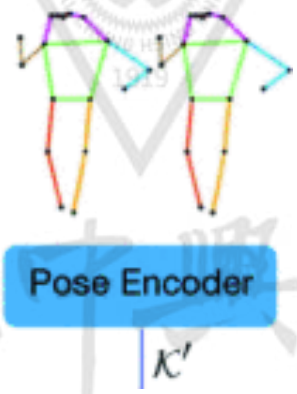


圖 8 單幀骨架編碼示意圖[9]

既有方法雖各有優勢，然而在桌球擊球辨識等高動態且細節豐富任務中，仍面臨挑戰。HIT Network 中的 Pose Branch 僅考慮單幀骨架特徵，缺乏對時序變化的建模能力，使其難以區分動作細節的微妙差異。另一方面，SkateFormer 在建模骨架與時間關聯上具備優勢，惟其單模態特性容易受到骨架遮蔽與估計誤差干擾。

為此，本研究延續 HIT Network 的雙模態架構，提出以 SkateFormer 取代原有 Pose Branch，導入具時序建模能力之骨架特徵學習模組，同時融合球拍幾何與外觀資訊，以提供骨架之外的輔助訊號。此設計旨在提升模型對關鍵動作語意之區辨能力，強化對複雜擊球動作的識別準確性與穩定性。

第三章 研究方法

本研究之整體流程如圖 9 所示，首先對輸入影片進行前處理，包含球拍區域的面積分割與人體姿態估測。我們採用物件分割模型進行球拍區域偵測與遮罩生成，並據以計算該區域之邊界框、面積與中心位置；同時，透過二維姿態估計模型偵測選手身體各關節點的空間座標。接著，將所取得之骨架資料進行時序特徵建構，輸入至骨架導向的深度網路中，提取時間與空間的運動語意。為提升辨識精度與語意理解，我們進一步整合骨架與 RGB 兩模態資訊，使用雙模態動作辨識架構，並引入球拍區域的幾何特徵 (面積與中心座標) 作為輔助線索以強化跨模態交互。最終，辨識結果將結合原始影像與各模態資訊，整合後同步顯示於輸出影片中，以利後續分析與可視化呈現。

本章節中，3.1 節將介紹球拍分割所使用之物件分割網路架構；3.2 節則說明人體姿態估測模型；3.3 節將深入探討骨架時序導向之動作識別架構；最後於 3.4 節介紹雙模態融合方法與交互模組設計。

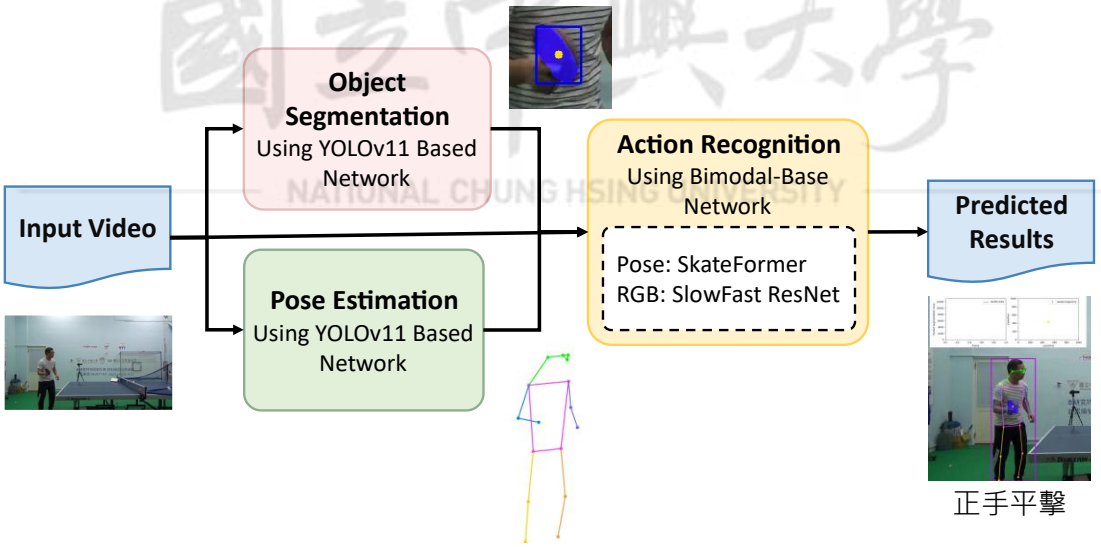


圖 9 所提方法之流程圖

3.1 物件分割模型 (Object Segmentation Model)

隨著人工智慧 (Artificial Intelligence, AI) 技術的迅速發展，深度學習 (Deep Learning) 已在眾多領域中展現強大效能，特別是在電腦視覺 (Computer Vision) 領域。透過深度學習模型，影像中的物件可以被自動偵測、分類，甚至像素級分割，使模型能夠理解影像中的空間與語意資訊，並應用於各類高層次視覺任務，如目標追蹤、動作辨識與場景解析等。

在這些應用中，卷積神經網路 (Convolutional Neural Network, CNN) 扮演核心角色。CNN 通常由特徵擷取模組與分類模組構成，其中前者由多層卷積層 (Convolutional Layer) 與池化層 (Pooling Layer) 組成，用以逐層提取輸入影像的不同層次特徵。卷積運算透過具參數化的卷積核 (Kernel)，在特定步長 (Stride) 下滑動並提取局部區域特徵，進而構建具備語意資訊的特徵圖 (Feature Map) 作為後續處理之基礎。最終，模型可依任務需求透過輸出層 (Output Layer) 進行分類，或結合分割頭 (Segmentation Head) 與姿態估計模組 (Pose Estimation Module) ，實現物件的像素級分割與人體關節點 (Keypoints) 推論。訓練過程中則藉由損失函數 (Loss Function) 評估模型預測與真實標註 (Ground Truth) 間的誤差，並利用反向傳播演算法 (Backpropagation) 更新網路權重，以最小化預測誤差、提升整體模型性能[10]。

在眾多影像處理技術中，物件分割 (Object Segmentation) 提供比傳統物件偵測更細緻的資訊。傳統偵測方法僅以邊界框 (Bounding Box) 粗略標示物件，而分割任務則能對每個像素進行分類，取得物件更精確的輪廓。語意分割方法如 Fully Convolutional Networks (FCN) 可以對每個像素進行類別預測，但是無法區分屬於同一類別的不同實體[11]。為了解決此問題，後續如 Mask R-CNN [12] 與 YOLACT [13] 等方法進一步結合偵測與像素分割機制，實現了可區分實體的即時實例分割能力 (Instance Segmentation) 。

桌球運動中，球拍具備快速移動、形狀狹長且外觀與背景容易混淆的特性，使其成為一個挑戰性的分割對象。針對此情境，實例分割技術提供了更高解析度的物件輪廓資訊，有助於準確掌握球拍的空間位置與形變狀態，進一步強化後續動作識別模型對於擊球動作的辨識效能。

為提升模型效能與應用即時性，現今主流的實例分割方法多採用一階段 (One-Stage) 架構，具備端到端 (End-to-End) 訓練與推論優勢。YOLO (You Only Look Once) 系列為其中代表，其設計在速度與準確度之間取得良好平衡。

YOLOv11 [14] 延續 YOLO 架構的高效性，並整合快速空間金字塔池化模組 (Spatial Pyramid Pooling - Fast, SPPF) [15] 與路徑聚合網路 (Path Aggregation Network, PANet) [16]，強化多尺度特徵圖的整合能力，提升模型對於尺寸變化大或邊緣模糊物件的辨識準確度。針對桌球球拍這類邊緣容易模糊，且外觀易與背景或選手身體部位混淆的物件，YOLOv11 能提供穩定且高精度的分割效果，作為後續動作辨識階段的重要輸入資訊。

在本研究中，我們採用圖 10 所示之 YOLOv11 分割架構，並以手工標註之球拍遮罩資料進行端對端訓練，訓練後的模型可針對每一影像框架輸出球拍的像素級遮罩結果，提供準確的區域輪廓資訊。

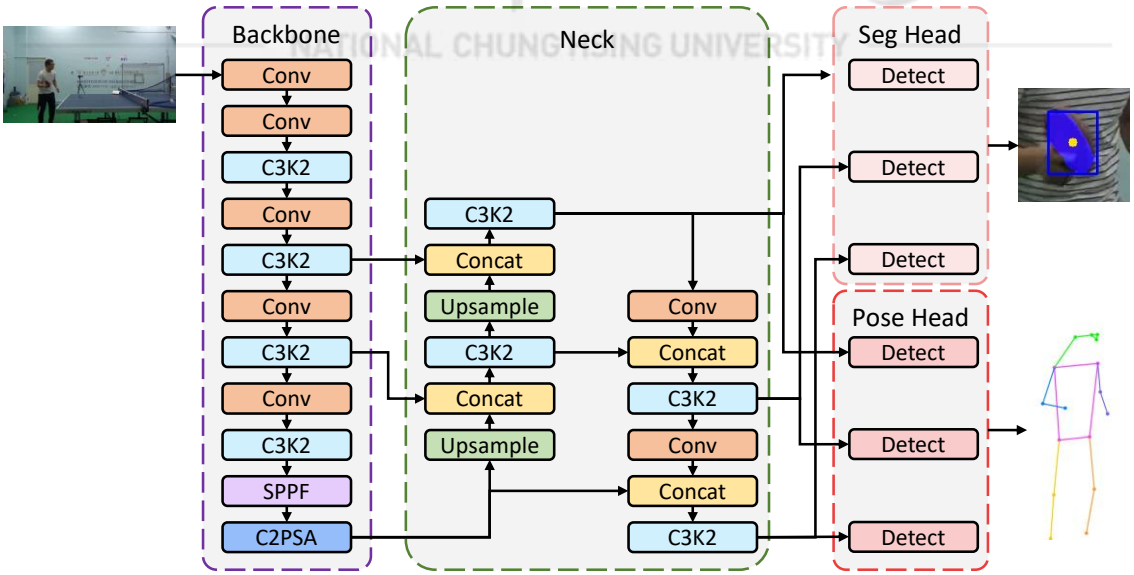


圖 10 物件分割暨人體姿態估計網路架構圖

當遮罩預測完成後，本研究進一步提取與球拍相關的多項幾何資訊，包括以遮罩區域計算之邊界框位置、像素面積與中心座標，作為後續動作識別模型之輸入特徵。這有助於更精準地掌握球拍在擊球過程中的運動軌跡與空間變化，進而提升細微動作辨識的準確性與穩定性。

3.1.1 快速空間金字塔池化模組 (Spatial Pyramid Pooling - Fast)

為提升模型在分割任務中的多尺度表徵能力，YOLOv11 [14] 採用了改良後的快速空間金字塔池化模組 (Spatial Pyramid Pooling - Fast, SPPF) [15]，其結構如圖 11 所示。該模組基於傳統的空間金字塔池化 (SPP) [17] 進行簡化設計，設計目的在於兼顧多尺度特徵擷取與推論效率。在原始 SPP 中，模型需要分別執行多個不同核大小 (例如 5、9、13) 的最大池化操作，並將結果拼接 (concatenate) 以取得不同感受野的特徵。然而這種設計會導致計算資源需求上升，尤其在高解析度輸入下更為明顯。

為改善此問題，SPPF 採用重複堆疊相同大小 (Kernel Size=5) 的最大池化層，模擬出類似多尺度感受野的效果。SPPF 將原始特徵圖經過一層卷積處理後，依序經過三層最大池化操作，每層輸出的特徵會回饋並與初始輸出進行串接處理。這種堆疊式設計，不僅能有效聚合局部與全域特徵，同時可以避免引入額外參數與大量運算成本，使得 SPPF 模組在保有表現力的同時，顯著提升整體推論效率。

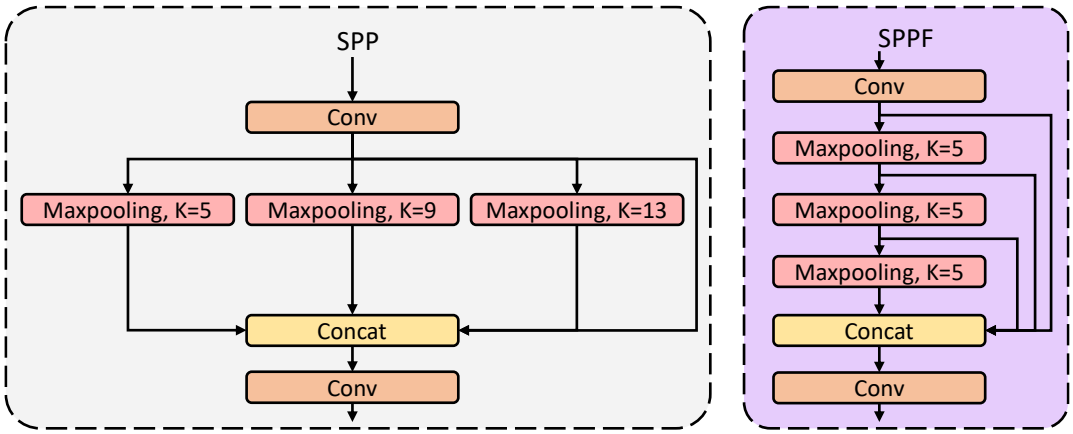


圖 11 SPP 與 SPPF 模組架構比較

針對本研究所關注的桌球球拍分割任務而言，球拍具有數個分割上的挑戰特性：其一為邊界模糊與形狀變異性高，尤其在快速揮拍時會產生運動模糊，使得球拍與背景或選手身體的邊緣難以區分；其二為尺寸變動劇烈，因為球拍在近拍與遠拍場景中的尺寸差異可能非常大；其三為類似背景干擾高，如球桌邊緣、衣物等，顏色與紋理皆可能與球拍部分區域高度相似。

面對上述挑戰，SPPF 所提供的多尺度特徵表徵能力正好發揮關鍵作用。不同層級的最大池化操作可捕捉從小尺度的邊界細節到大尺度的整體形狀輪廓，幫助模型在模糊或邊界不清晰時仍能保有對球拍形狀的正確理解。此外，堆疊結構使得特徵具備層層整合的上下文語意，有助於判斷是否為「真實球拍」區域，而非錯誤分割的類似背景物件。

3.1.2 路徑聚合網路 (Path Aggregation Network)

為進一步強化特徵表徵能力與上下層級間的訊息流動，YOLOv11 [14] 採用了路徑聚合網路 (Path Aggregation Network, PANet) [16] 作為其頸部網路 (Neck) 結構，以提升模型於不同層級特徵之間的整合能力，進而強化多尺度特徵圖之間的資訊傳遞與語意融合。此設計理念源自 Lin 等人所提出之特徵金字塔網路 (Feature Pyramid Network, FPN) [18]，如圖 12 所示。

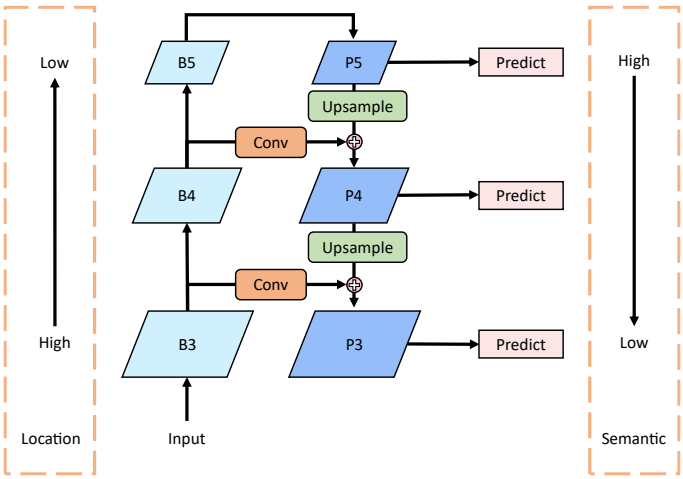


圖 12 特徵金字塔架構圖

FPN 架構藉由自頂向下 (top-down) 的語意傳遞流程，將深層特徵所蘊含之語意資訊傳播至淺層特徵圖，並與其保留之空間細節進行融合，有效提升模型於多尺度物件辨識中的整體效能。FPN 解決了傳統卷積神經網路中，語意深度與空間解析度之間的落差問題，使模型能兼具語意判別與空間定位的能力。然而，FPN 僅提供單向之語意傳遞，底層特徵中所蘊含的豐富細節難以向上傳遞至高層進行補強，導致語意一致性與邊界辨識的穩定性可能受限。

為克服此限制，Liu 等人提出 PANet 架構 [16]，於 FPN 基礎上加入自底向上 (bottom-up path augmentation) 之增強路徑，藉由跳接 (skip connection) 與上採樣 (upsampling) 機制，將低層特徵回饋至中高層級進行語意修正與表徵補強。最終形成一套具備雙向資訊流的強化特徵融合架構，如圖 13 所示，不僅保留了 FPN 所提供之語意下傳能力，更強化了特徵間的語意一致性與空間完整性。

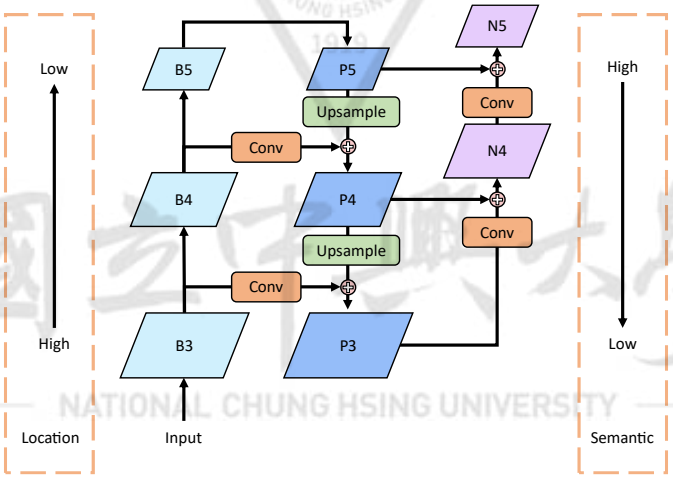


圖 13 路徑聚合網路架構圖

相較於 SPPF 所著重的單層內部之多尺度特徵強化策略，PANet 更進一步處理跨層級特徵融合問題，並針對語意與空間資訊的交互補足提出具體解法。球拍屬於中小尺寸、狹長且運動快速之物件，若僅依賴高層語意特徵進行分割，常因解析度不足而造成邊界遺失；若僅依賴低層特徵，則雖具定位資訊卻缺乏語意脈絡。PANet 所建構之雙向特徵融合路徑，能將高層語意傳遞至底層以輔助精確定位，並將底層細節回饋至高層以進行語意修正，使模型能維持穩定且精準之分割效能。

3.2 二維人體姿態估計模型 (2D Human Pose Estimation Model)

二維人體姿態估計 (2D Human Pose Estimation) 旨在從靜態影像中推斷人體的姿勢結構，透過預測人體各部位的關節點 (Human Body Keypoints)，如肩膀、手肘、髖部等，建立具備幾何關係的骨架資訊，如圖 14 所示。此任務通常包含關節點定位、關節配對與姿態優化等步驟，目的是還原人物在影像中的真實姿態。

圖 14 二維人體姿態估計示意圖

在桌球運動場景中，許多擊球動作彼此極為相似，細微差異往往體現在關節角度的變化與揮拍動作的節奏之中。透過精準的姿態估計，不僅能有效捕捉這些細節變化，也有助於提升後續動作辨識的準確性。

本研究採用的 YOLOv11 [14] 架構如圖 10 所示，屬於一種單階段 (single-stage) 姿態估計模型，可於進行人物偵測的同時，針對每一個檢測框 (bounding box) 同步預測 17 個人體關節點位置。相較於傳統兩階段架構需額外套用姿態子模型，YOLOv11 在單一前向傳遞中即可完成人物位置與骨架資訊之預測，具備高效率與即時性。其關節點回歸機制基於多層特徵圖，融合來自不同解析度與語意層級的訊息，使模型得以同時兼顧整體人體結構與細部肢體動作之解析。

為強化骨架估計於背景雜訊、遮擋與模糊情境下的表現，YOLOv11 架構中引入多項高效模組。C3K2 透過堆疊式 3×3 卷積強化對關節邊界與紋理的辨識；C2PSA 則以平行空間注意力機制 (Parallel Spatial Attention) 提升關鍵區域的感知能力。此外，本研究採用官方釋出之 YOLOv11-pose 預訓練權重以加速實驗流程。

3.2.1 雙層 3x3 卷積交叉階段模組 (Cross Stage Partial with two 3 x3 convolution layers)

YOLOv11 [14] 架構於特徵提取階段中採用數種進階模組以提升對人體關節點之辨識能力。其中特徵提取主幹 (backbone) 延續了 YOLOv8 [19] 中所引入之 C2F (Cross Stage Partial with Two-Branch Fusion) 模組，並進一步導入強化版的 C3K 與 C3K2 結構，如圖 15 所示。這些模組在保持運算效率的同時，提升模型對於細節特徵的捕捉能力，特別適用於辨識動作關鍵區域如關節、肢幹交界等位置。

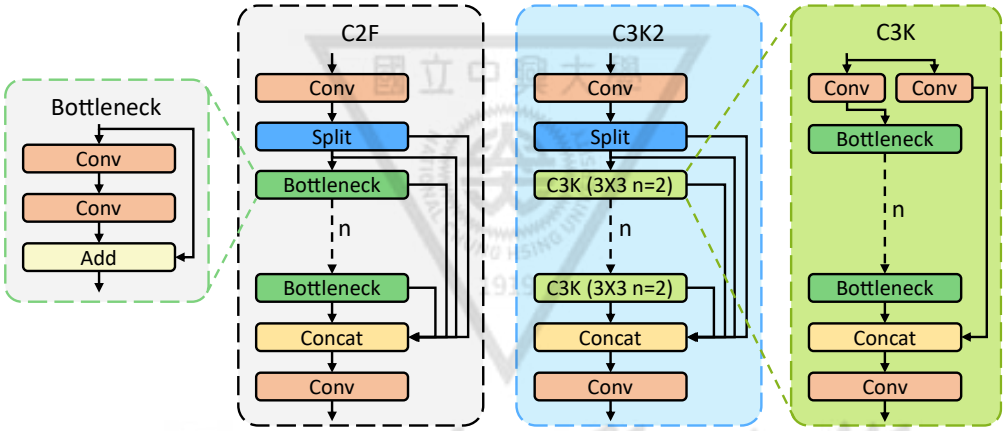


圖 15 C2F、C3K2 與 C3K 模組結構比較示意圖

C2F 模組為 CSPNet 架構 [20] 之延伸，其設計核心在於將輸入特徵圖分為兩支路徑，部分特徵直接繞過主路徑進行跨層融合，避免重複冗餘計算，並促進梯度流傳遞。在主路徑中，模組透過多層 Bottleneck 結構進行深層特徵提取，最後再與繞過分支進行融合。這樣的架構設計有助於平衡特徵重用與計算負擔，使模型能夠在保留低階空間資訊的同時，擷取高階語意特徵。

YOLOv11 架構於其主幹網路中整合了 C3K2 模組作為核心運算單元，以強化特徵提取階段的空間表徵能力與深層特徵建模效能。C3K2 模組源自先前提出之 C3 結構 [15]，其主要設計差異在於內部固定包含兩層 Bottleneck 單元，並於每層均採用 3×3 卷積核進行堆疊。此設計不僅提高了特徵轉換的深度，亦增強了模組的非線性建模能力與空間感知精度。

與傳統 Bottleneck 結構常搭配 1×1 卷積進行通道壓縮與特徵融合不同，C3K2 模組以 3×3 卷積構成路徑，提升模型對邊界細節與局部紋理變化的感知敏銳度。此特性使其在面對如手肘、膝蓋等高變異關節點位置時，能提供更穩定的空間回歸效果。此外，雙層堆疊設計也促進了梯度訊息的深層傳遞與多層次特徵的融合，有助於提升模型於姿態遮擋與動作劇烈變化等挑戰情境下的辨識魯棒性。

3.2.2 平行空間注意力卷積模組 (Convolutional block with Parallel Spatial Attention)

YOLOv11 [14] 架構中引入了平行空間注意力模組 (Cross Stage Partial with Parallel Spatial Attention, C2PSA) ，其結構如圖 16 所示。該模組結合了 CSPNet [20] 設計理念與空間注意力機制，旨在於保持計算效率的前提下，強化模型對於關鍵區域的感知能力。

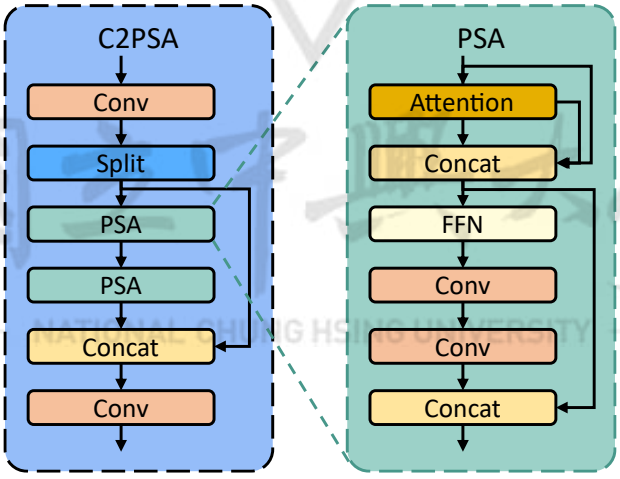


圖 16 平行空間注意力模組 (C2PSA) 與 PSA 子模組結構示意圖

C2PSA 模組首先透過卷積層進行初步特徵轉換，接著將輸入特徵圖進行通道分割 (Split)，並分別送入兩個 PSA (Parallel Spatial Attention) 子模組進行獨立處理。每個 PSA 子模組內部包含 Attention、Feed Forward Network (FFN)、兩層卷積與特徵串接操作，透過殘差設計保留輸入資訊，實現對空間特徵的選擇性強化。最終，兩條路徑所提取之注意力特徵將進行串接，再經由卷積整合輸出。

此模組設計可有效捕捉局部與全域空間脈絡，有助於模型在雜訊干擾、背景複雜的情境中，聚焦於具有辨識意義之關鍵部位。對於二維人體姿態估計任務而言，C2PSA 尤其適用於關節點區域特徵極為細緻、對比度低或與背景混淆度高的情況。其多層次空間注意力聚焦機制能提升模型對如手肘、手腕、腳踝等部位的感知準確性，進而強化姿態預測的整體穩定性與可靠性。

3.3 骨架時序導向動作識別模型 (Temporal Skeleton-Based Action Recognition Model)

骨架資訊具備高度結構性與動作語意表達能力，能有效排除背景干擾與外觀變異，特別適合應用於人體動作的建模與辨識任務。為提升模型對動作時序連續性與跨關節依賴關係的捕捉能力，本研究採用 SkateFormer 作為骨架模態之主幹模型，其整體架構如圖 17 所示。該模型融合時間序列建模與骨架結構感知能力，能同時考量動作發展的時序變化與關節間的空間關聯性，並透過多層次特徵抽取與資訊壓縮機制，有效強化模型在處理複雜時空模式與細節動作變化上的表現，具備良好的辨識穩定性與語意理解能力。

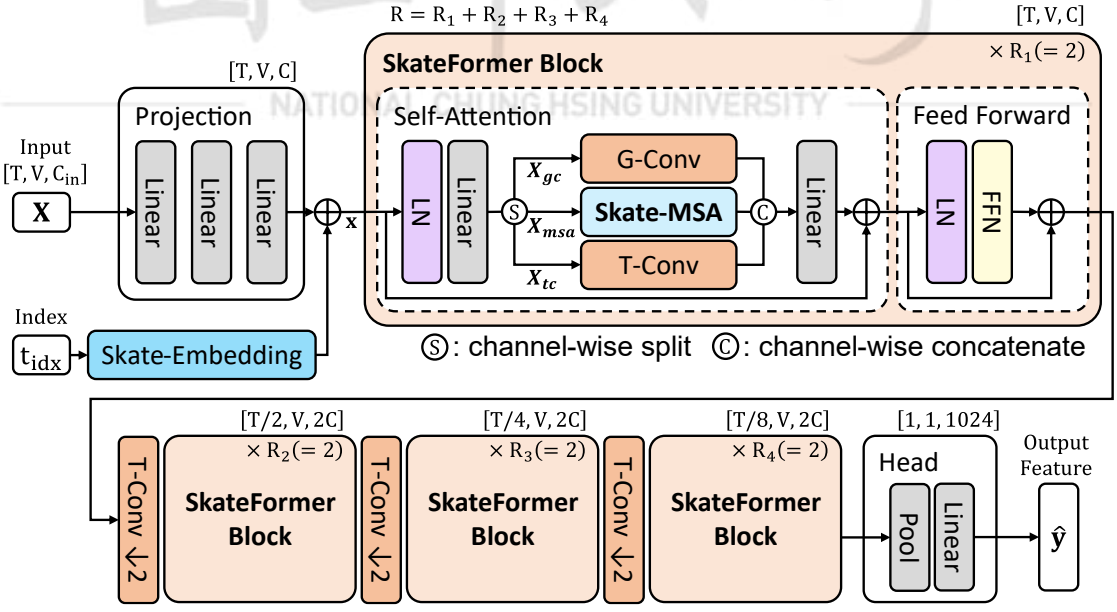


圖 17 SkateFormer 架構圖

SkateFormer 的輸入為一段長度為 T 的骨架序列 $X \in \mathbb{R}^{T\times V\times C_{in}}$，其中 V 每一幀的關節點數，C_{in} 表示每個關節的維度資訊。首先，輸入資料會透過三層線性投影模組 (Projection Layer) 將原始低維的骨架資料映射至高維特徵空間，接著加入骨架－時間位置嵌入 (Skate-Embedding)，增強模型對時序與骨架結構的理解。

特徵經過加權後進入主體 SkateFormer Block，其中每個 Block 採用通道分割設計的 Self-Attention 模組組成，包含骨架圖卷積 (G-Conv)、時間卷積 (T-Conv) 與骨架多頭自注意力模組 Skate-MSA (Skeletal-Temporal Multi-Head Self-Attention)，能從骨架間與時間軸兩個方向捕捉跨關節與時序依賴。每個 Block 中皆搭配前饋神經網路 (Feed-Forward Network, FFN) 與殘差結構，以提升特徵穩定性與表達力。

整體網路堆疊四個階段的 SkateFormer Block，分別對應時間維度下採樣的四個層級T、$\frac{T}{2}$、$\frac{T}{4}$、$\frac{T}{8}$，每個層級均搭配 2 次重複堆疊$(R_1 = R_2 = R_3 = R_4 = 2)$，最後的輸出特徵將沿時間與節點維度進行聚合 (Pooling)，並接續一層線性變換，轉換為固定維度的骨架語意向量，用作後續交互模組之輸入。

3.3.1 骨架與時間關係建模 (Modeling Skeletal and Temporal Relations)

在以骨架為基礎的動作辨識任務中，骨架關節 (joints) 於時間軸上的連續排列構成了動作語意的重要來源。傳統方法多採用圖卷積網路 (GCNs) 以建模關節間之鄰接關係，然而由於其訊息傳遞通常侷限於物理相鄰的節點，對於如手腕與肩膀之間這類遠距但具語意關聯的互動關係建模能力有限。此外，在時間建模方面，此類方法普遍仰賴固定視窗的一維時間卷積進行序列處理，僅能捕捉相鄰幀之間的局部變化，無法有效涵蓋整個動作從起始到結束的長期語意脈絡。例如從預備動作、施力階段至擊球後的收尾動作，往往涉及跨時間段的資訊整合，而傳統時間卷積難以捕捉此類動作演化過程的完整結構與節奏特徵。

為克服此限制，SkateFormer 採用基於 Transformer [8] 的結構，透過自注意力機制 (Self-Attention) 整合關節之間在空間與時間兩個面向的互動，並提出結合骨架與時序的分區式建模策略，進一步提升對複雜動作語境的理解能力，使模型能更加靈活地捕捉長距離的時空依賴與跨部位的語意聯繫。

為實現前述的分區式建模策略，本研究首先將整個人體關節集依據其空間位置與功能特性，劃分為完全不重疊的 K 個子集合，作為相鄰關節分區 (Neighboring Joint Partitions)，記為 $\{v_k^{njp}\}_{k=1}^{K}$。當 $K=8$ 時，分區 $\{v_k^{njp}\}_{k=1}^{8}$ 可依序表示為 v_1^{njp}、v_2^{njp}、v_3^{njp}、v_4^{njp}、v_5^{njp}、v_6^{njp}、v_7^{njp}、v_8^{njp}，分別對應右手臂、左手臂、右腿部、左腿部、肩膀、腰部、右眼耳與左眼耳等八大分區，如圖 18 所示。

每個子集合內的關節節點則按照由身體中心向外延伸的順序進行排列，反映實際動作過程中的空間連續性與運動傳遞路徑。透過此種區塊化設計，模型能針對各部位的語意特徵與空間互動進行獨立建模，有助於提升特徵學習效率與辨識準確性，並進一步強化對不同動作模式下關鍵部位變化的捕捉能力。

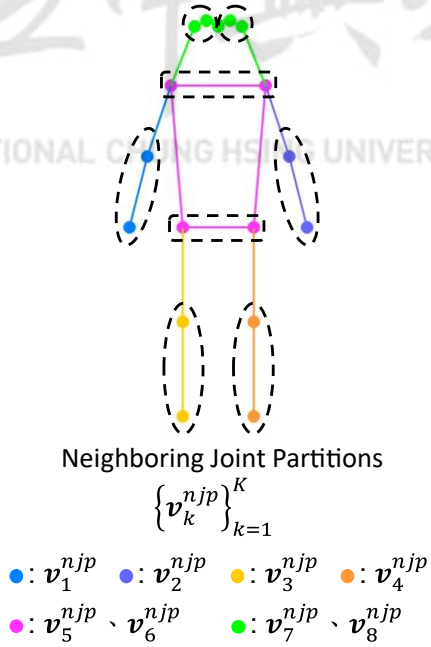


圖 18 人體骨架分區示意圖

此外，如圖 19 所示，SkateFormer 進一步將骨架－時間關係劃分為兩大面向：空間上的關節關係 (Skeletal Relation Types) 與時間上的幀間關係 (Temporal Relation Types)，並依據關節相對位置與時間距離設計出四種注意力策略，用以指引模型關注最具判別力的語意組合。

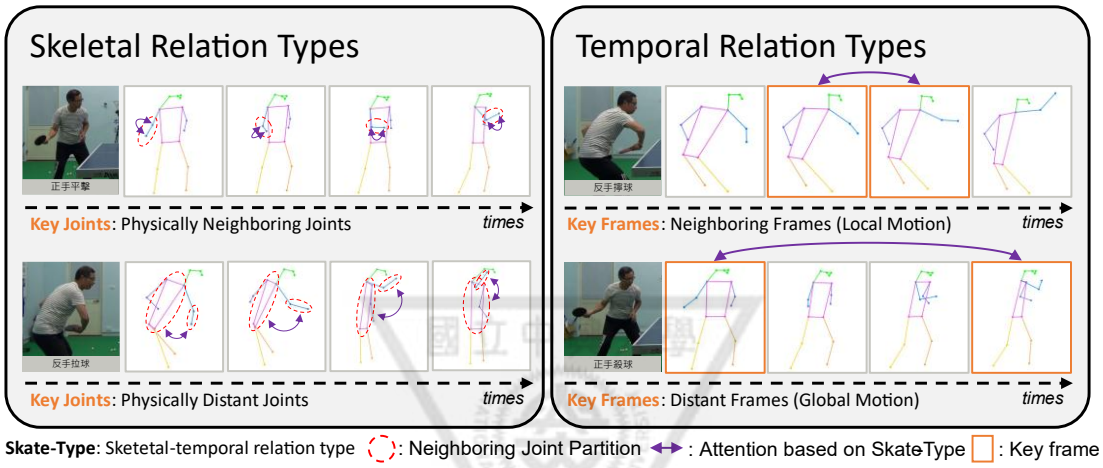


圖 19 骨架與時間關係的類型示意圖

在空間面向，模型區分「鄰近關節」與「遠距關節」：前者對應如正手平擊揮拍動作中手腕與手肘之間的連續變化，能捕捉區域內的關節細節；後者則如反手拉球手臂動作帶動肩膀旋轉時所出現的長距離關節互動，對於建構全身協調性動作的語意極為重要。

在時間面向，模型區分「局部時間」與「全域時間」：前者對應如反手擰球等短時間內的細微手軸調整，能捕捉動作瞬間的細節變化；後者則如正手殺球這類需要觀察較長時間範圍內動作發展，對於建構整體動作結構與節奏具有關鍵作用。

為有效捕捉上述語意脈絡，SkateFormer 提升對不同關節與時序關係的區辨能力，並能自動從動作序列中辨識並加權關鍵幀 (key frames) 與關鍵關節 (key joints)，使模型聚焦於最具判別性的時空位置，進一步提升語意表徵能力。此策略同時克服 GCN 難以建模遠距關節的限制，並緩解傳統 Transformer 所帶來的高計算成本與資訊稀釋問題，提升辨識效率與準確性。

3.3.2 骨架時序轉換器 (Skeletal-Temporal Transformer)

SkateFormer Block 是整體架構的核心組件，其設計承襲傳統 Transformer [8] 的架構，包含自注意力層 (Self-Attention Layer) 與前饋神經網路 (Feed-Forward Network, FFN)。

如圖 20 所示，為了同時捕捉骨架動作的空間拓撲與時間演化特性，SkateFormer 將輸入特徵透過層正規化 (Layer Normalization) 和線性層處理後，依照通道維度進行三路分割，分別對應骨架圖卷積 (G-Conv)、時間卷積 (T-Conv) 與骨架多頭自注意力模組 (Skate-MSA)。其中，輸入特徵通道被拆分為$x_{gc} \in \mathbb{R}^{C/4}$、$x_{tc} \in \mathbb{R}^{C/4}$、$x_{msa} \in \mathbb{R}^{C/2}$三個子分支，分別聚焦於空間、時間與語意上的資訊抽取。

在空間層面，G-Conv 模組引入一個形狀為 $(H/4, V, V)$ 的可學習關節關聯矩陣，取代預先定義的靜態鄰接圖，使每個分支能建構不同的潛在連接模式，提升模型對關節結構關係的表達能力。在時間層面，T-Conv 採用核大小為 k 的一維卷積操作，針對每組 $(H/4)$ 通道的特徵進行時間序列建模，能有效建構關節隨時間推移的局部變化趨勢，特別適合解析短時動作的細微動態。而 Skate-MSA 則設計為具感知的分區式自注意力機制，可依據關節距離與時間跨度之組合，自動針對不同的關節－時間關係型態進行加權建模。

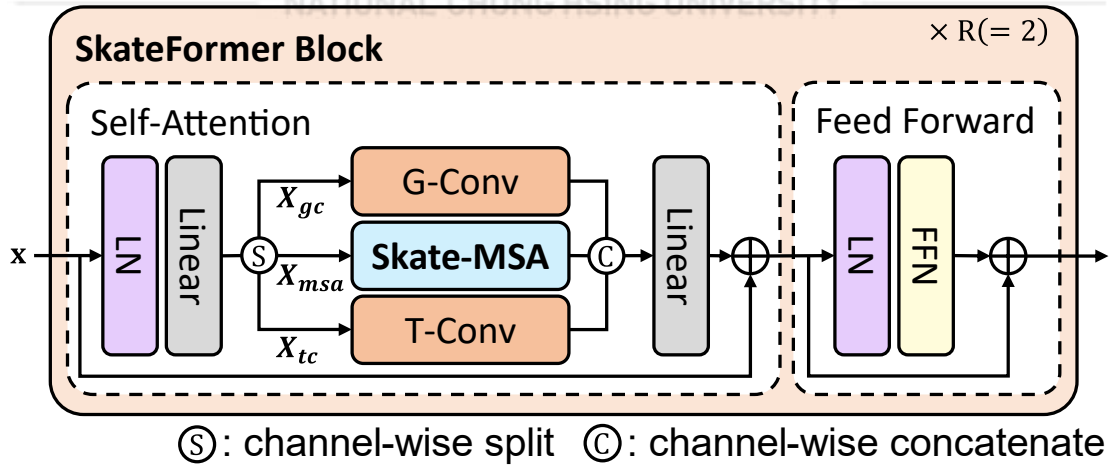


圖 20 SkateFormer Block 模組架構示意圖

三個子分支處理完後，SkateFormer Block 會將 G-Conv、T-Conv 與 Skate-MSA 的輸出沿通道維度進行串接 (concatenate)，形成整合後的特徵表示，接著通過線性層進行維度映射，融合三者資訊後再與原始輸入特徵進行殘差連結 (residual connection)，以完成自注意力階段的特徵更新。此步驟不僅有助於保留輸入訊號中的關鍵資訊，也能提升深層網路在訓練過程中的穩定性。上述運算流程可形式化表示如下式(1)表示：

$$
\begin{aligned}
[x_{gc}, x_{tc}, x_{msa}] &= \text{Split}\Big(\text{Linear}\big(\text{LN}(x)\big)\Big) \\
x_{gc} &\leftarrow \text{G-Conv}\big(x_{gc}\big) \\
x_{tc} &\leftarrow \text{T-Conv}(x_{tc}) \\
x_{msa} &\leftarrow \text{Skate-MSA}(x_{msa}) \\
x &\leftarrow x + \text{Linear}\Big(\text{Concat}\big(x_{gc}, x_{tc}, x_{msa}\big)\Big)
\end{aligned}
\tag{1}
$$

在完成自注意力階段後，SkateFormer Block 還包含一個前饋神經網路 (Feed-Forward Network, FFN)，該模組由兩層線性變換和中間的 GELU 激勵函數所組成，並搭配殘差連結與層正規化 (Layer Normalization) 以強化非線性表達能力與收斂穩定性。此設計延續傳統 Transformer [8] 架構的優勢，能有效提升特徵轉換過程中的深層語意建模能力。其運算流程可表達如下式(2)表示：

$$
x \leftarrow x + \text{Linear}\Big(\text{Act}\Big(\text{Linear}\big(\text{LN}(x)\big)\Big)\Big) \tag{2}
$$

為提升模型對長時間序列的理解，模型在部分區塊開頭加入 $\text{stride} = 2$ 的一維卷積與 Batch Normalization 組成的下採樣模組，用於保留語意的同時壓縮時間長度，擴展時間感受野。SkateFormer Block 結合空間層的 G-Conv、時間層的 T-Conv 與語意層的 Skate-MSA 模組，搭配殘差連結、特徵融合與下採樣設計，實現骨架－時間－語意的多層次建模，能有效捕捉跨關節互動、時間演化趨勢與語意關聯，進一步強化骨架時序建模能力與動作辨識精度。

3.3.3 骨架多頭自注意力模組 Skate-MSA (Skeletal-Temporal Multi-Head Self-Attention)

為進一步強化模型對骨架時序語意的辨識能力，SkateFormer 在每個區塊中導入專為骨架設計的多頭自注意力模組 (Skate-MSA)。所圖 21 示，Skate-MSA 模組首先將輸入特徵x_{msa}沿通道維度均分為四組子特徵x_{msa}^1、x_{msa}^2、x_{msa}^3、x_{msa}^4，每組通道數為$\frac{C}{8}$。這四個分支分別對應於四種骨架與時間關係 (鄰近關節與局部時間、遠距關節與局部時間、鄰近關節與全域時間、遠距關節與全域時間)，並且每一分支特徵會進行對應的切割操作 $P_i(\cdot)$，將原始骨架－時間序列重新排列為適合該類關係的結構格式，進而強化模型對特定空間與時間區塊的注意力學習。切割後的張量輸入至多頭自注意力 (Multi-Head Self-Attention, MSA) 模組中進行語意關聯建模，隨後透過還原操作 $R_i(\cdot)$，將注意力輸出結果還原回原始位置。整體運算流程可形式化如式(3)表示：

$$x_{msa}^i \leftarrow R_i\left(\text{MSA}\left(P_i\left(x_{msa}^i\right)\right)\right), \qquad for\ i = 1,2,3,4 \tag{3}$$

最後，四個處理後的子特徵向量將以通道串接 (channel-wise concatenation) 方式進行整合，形成最終語意特徵x_{msa}，如式(4)表示：

$$x_{msa} \leftarrow \text{Concat}(x_{msa}^1, x_{msa}^2, x_{msa}^3, x_{msa}^4) \tag{4}$$

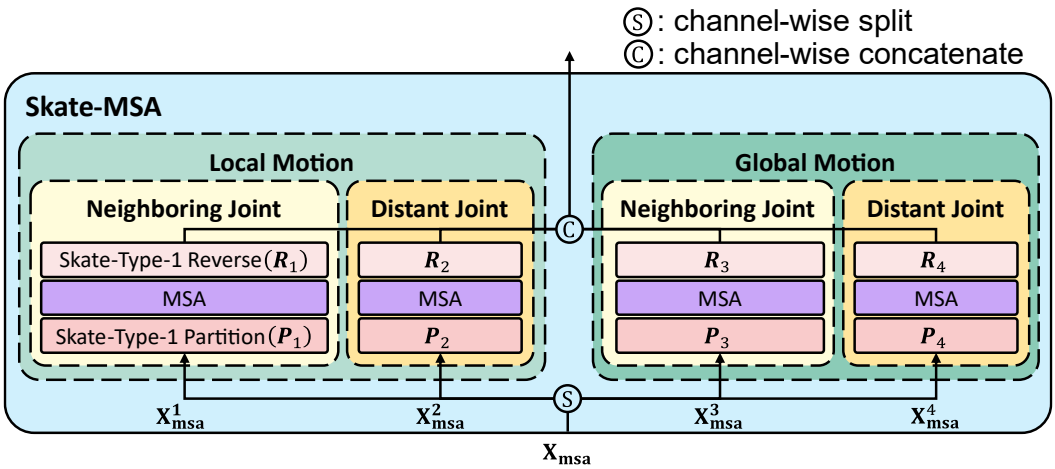


圖 21 Skate-MSA：四分支特徵依據空間與時間關係分割與還原後整合

3.3.4 骨架與時序的分區與反轉 (Skeletal and Temporal Partitioning and Reversal)

為精準建模人體動作中的語意關聯與時間結構，Skate-MSA 採用一套結合空間與時間的分區 (Partitioning) 與還原 (Reversal) 策略，針對不同的骨架－時間關係進行特徵重組，如圖 22 所示。該策略依據關節的空間鄰近性與時間連續性，將輸入特徵切割成具結構與時間對齊性的區塊，使每個注意力分支能專注於特定的骨架分區與時間範圍，強化對關節互動與動作時序邏輯的建模能力。透過將原始骨架序列拆解為語意一致的小分區，模型可在每一子空間內進行更細緻的特徵提取，並聚焦於具代表性的動作模式與區域互動。

此設計不僅提升模型對姿態的辨識效果，也有助於捕捉如快速切換、方向改變與肢體協調等動作細節。與此同時，還原操作可將切割後的特徵結果映射回原始骨架順序，維持時間與空間語意的一致性。該策略有效緩解傳統 Transformer [8] 在處理高維輸入時所面臨的資訊稀釋、注意力分散與計算成本高昂等問題，進而在動作辨識任務中展現更穩定且具判別力的表現。

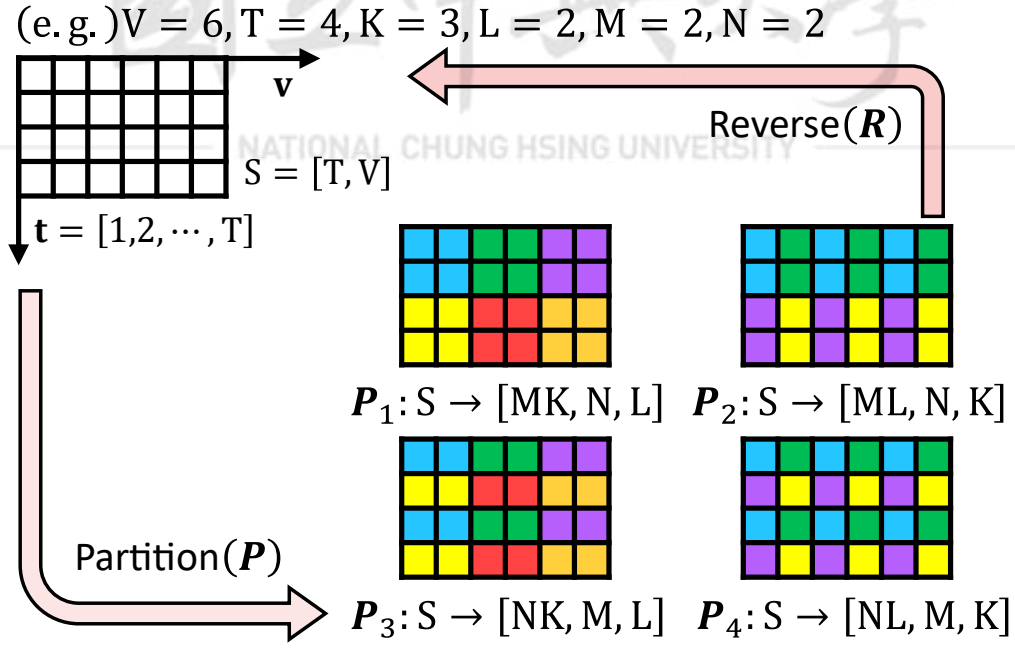
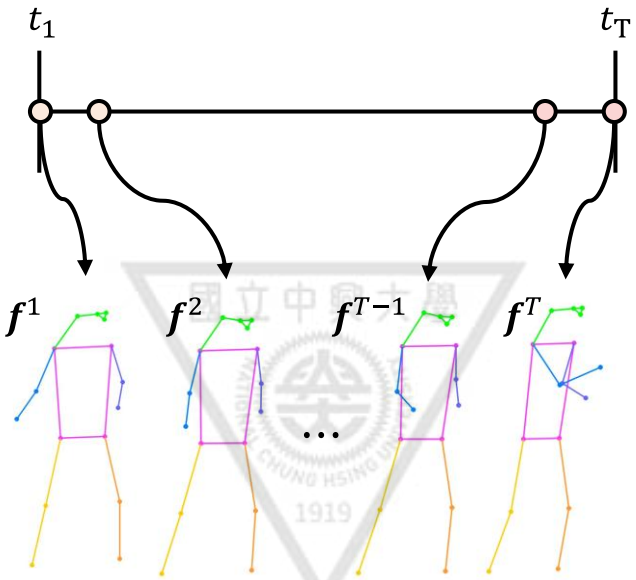


圖 22 骨架與時間特徵之結構重排與還原操作示意圖

如圖 23 所示，在每個時間點 t，一張骨架幀可表示為 $f^t \in \mathbb{R}^{V \times C_{in}}$，其中 V 為關節數，$C_{in} = 2$ 表示每個關節的 2D 座標。整段骨架序列由 T 幀組成，可表示如式(5)：

$$\mathbf{X} = \{f_t\}_{t=1}^{T} \in \mathbb{R}^{T \times V \times C_{in}} \tag{5}$$



$$\boldsymbol{X} = \{\boldsymbol{f}^t\}_{t=1}^{T} \in \mathbb{R}^{T \times V \times C_{in}}$$
$$t_{idx} = [t_1, t_2, \cdots, t_T]$$

圖 23 骨架序列輸入示意圖

在空間維度上，為建構動作中關節間的語意與結構關聯，Skate-MSA 將骨架中的關節依據不同策略劃分為兩種結構單元：鄰近關節分區 (Neighboring Joint Partitions, v_k^{njp}) 與遠距關節分區 (Distant Joint Partitions, v_l^{djp})，分別用以捕捉空間上鄰近部位與物理距離較遠但語意密切的關節互動。

鄰近關節分區依據身體解剖結構，將骨架中 V 個關節劃分為 K 個不重疊的分區，每個分區由若干具空間鄰近性的關節構成。以 $K = 8$ 為例，這些分區可依序對應為右手臂、左手臂、右腿部、左腿部、肩膀、腰部、右眼耳與左眼耳等八大部位。每個分區可形式化表示如式(6)：

$$v_k^{njp} = [v_{k,1}, v_{k,2}, \cdots, v_{k,L}], \qquad k = 1,2,\cdots,K \tag{6}$$

其中，L 表示每個鄰近關節分區中所包含的關節數。此外，所有分區彼此不重疊，亦即 $v_i^{njp} \cap v_j^{njp} = \emptyset, for\ i \neq j$。各分區內的關節排列方式以身體中心向外延伸為原則，以保留動作傳遞的方向性與語意一致性。例如， $k = 1$ 表示右手臂分區，則 $v_{1,1}$、$v_{1,2}$ 分別對應於右手肘與右手腕的 2D 座標；又如 $k = 3$ 為右腿部，則關節依序為右膝與右腳踝。

整體骨架可視為將所有鄰近關節分區沿維度串接後所形成的骨架軸，表示如式(7)：

$$v = \left[v_1^{njp} \middle| v_2^{njp} \middle| \cdots \middle| v_k^{njp}\right] \tag{7}$$

其中，$|v| = K \cdot L = V$，表示串接後的總關節數為 V。

遠距關節分區是為進一步建模遠距關節間的語意互動 (例如肩膀與對側手腕的協同動作) 所提出的結構單元。為此，模型將所有鄰近關節分區中相同位置的關節元素進行對齊組合，具體作法是對每一位置索引 l，自每個鄰近分區中擷取該位置的關節，並組合如式(8)：

$$v_l^{djp} = \left[v_{1,l}, v_{2,l}, \cdots, v_{K,l}\right], \quad l = 1,2,\cdots,L \tag{8}$$

其中，L 表示每個鄰近分區內的關節數。透過此對位重組策略，模型得以顯式建模物理距離雖遠、但語意上密切相關的關節互動，進一步提升對複雜動作結構的辨識能力。

在時間維度上，Skate-MSA 為更細緻地建模動作的階段性結構與動作幅度變化，將完整的時間軸 $t = [1,2,\cdots,T]$ 進一步劃分為兩種語意時間軸，分別為局部時間軸 (Local Motion, t_m^{local}) 與全域時間軸 (Global Motion, t_n^{global})，用以捕捉短期與長期的動作結構。

局部時間軸定義為長度為 N 的連續時間段，共有 $M = T/N$ 段，用以捕捉如 flick 動作中手腕等短時區域性動作，形式化定義如式(9)：

$$t_m^{local} = [(m-1)N + 1, (m-1)N + 2, \cdots, mN], \quad m = 1,2,\cdots,M \tag{9}$$

全域時間軸則透過每隔 N 步進行一次取樣，總共取得 N 條長度為 M 的片段，形成跨時間段的稀疏表示，適用於如 smash 等涵蓋全身運動的整體性動作，定義如式(10)：

$$t_n^{global} = [n, n+N, \cdots, n+(M-1)N], \qquad n = 1,2,\cdots,N \tag{10}$$

整體而言，這兩種語意時間軸可共同涵蓋完整的時間序列 t，其中局部時間捕捉短時動作變化，全域時間則對應長期結構，滿足如下關係式(11)：

$$t = \{t_m^{local}\}_{m=1}^{M} = \{t_n^{global}\}_{n=1}^{N}, \qquad |t| = M \cdot N = T \tag{11}$$

結合前述空間與時間的語意劃分，Skate-MSA 將骨架－時間序列依據不同關節與時間組合關係建構為四種結構類型，並分別定義為四組張量重組操作 $P_i(\cdot)$，其中 $i \in \{1,2,3,4\}$。

P_1是將鄰近關節分區 v_k^{njp} 與局部時間軸 t_m^{local} 結合，聚焦於局部動作中空間鄰近關節間的時序關聯。P_2是將遠距關節分區 v_l^{djp} 與局部時間軸 t_m^{local} 結合，用以強化短時動作中物理距離較遠之關節的語意互動。P_3是將鄰近關節分區 v_k^{njp} 與全域時間軸 t_n^{global} 結合，強調於長期動作中建構空間局部性結構。P_4是將遠距關節分區 v_l^{djp} 與全域時間軸 t_n^{global} 結合，捕捉全身性動作中的遠距關節協調。

上述四種張量重組操作會將輸入特徵張量$S \in \mathbb{R}^{T\times V\times c}$ (其中 T 為時間長度，V 為關節數，c 為通道數) 重排如式(12)的四種形式：

$$\begin{aligned} P_1 &: S \to [MK, N, L, c] \\ P_2 &: S \to [ML, N, K, c] \\ P_3 &: S \to [NK, M, L, c] \\ P_4 &: S \to [NL, M, K, c] \end{aligned} \tag{12}$$

每個重組後的張量皆會輸入對應的多頭自注意力模組 (Multi-Head Self-Attention, MSA)，以建模不同骨架－時間序列下的關聯，最終透過逆操作 $R_i(\cdot)$ 還原為原始形狀 (T, V, c)，以利後續通道維度上的拼接與融合。

3.3.5 多頭自注意力模組 MSA (Multi-Head Self-Attention)

當前述張量重組操作後，為了統一處理來自四種骨架－時間分區的語意資訊，模型首先將 Skate-MSA 中每一種分區類型所產生的特徵張量 $x_{msa}^{i,P}$ 調整為統一形狀 (B, T', V', c)，其中 B 為批次大小，T' 和 V' 分別表示時間軸與骨架軸在分區後的長度，c 則是通道數。接下來，這些特徵張量接著會被展平成 $(B, T' \cdot V', c)$ 的形狀，並進一步透過三組線性映射權重 $W_Q, W_K, W_V \in \mathbb{R}^{c \times c}$，分別轉換為查詢 $(\boldsymbol{Q})$ 、鍵 $(\boldsymbol{K})$ 、和值 $(\boldsymbol{V})$ 三個張量，其對應運算如式(13)：

$$[\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}] = x_{msa}^{i,P}[W_Q, W_K, W_V] \tag{13}$$

整體模型共包含 H 個注意力頭，其中一半 (即 $H/2$) 分配給 G-Conv 與 T-Conv 分支，剩餘的一半平均分配給四種分區類型，每一分區類型對應 $H' = H/8$ 個注意力頭。對於每個注意力頭 h，其查詢、鍵與值張量會被進一步 reshape 為 $(B, H', T' \cdot V', c/H')$，並執行標準的自注意力操作(Self-Attention, SA)，如式(14)：

$$SA_h(x_{msa}^{i,P}) = \text{SoftMax}\left(\boldsymbol{Q}_h \boldsymbol{K}_h^{tr} / \sqrt{c/H'} + \boldsymbol{B}_h\right) \boldsymbol{V}_h \tag{14}$$

其中，$\boldsymbol{Q}_h \boldsymbol{K}_h^{\text{tr}}$ 是一項與輸入特徵相關的資料依賴項，會隨著每個輸入樣本而變化，用以反映當前特徵間的相似性。而 $\boldsymbol{B}_h$ 為結構性位置偏差項，提供固定的語意指引，使注意力計算不僅依賴資料內容，亦能考慮時間與骨架空間中的相對結構。

在時間軸方面，Skate-MSA 採用一維相對位置偏差 $\boldsymbol{B}_h^t \in \mathbb{R}^{T' \times T'}$，以捕捉局部與全域時間軸的相對關係。而在骨架軸上，根據分區方式進行不同建模：對於鄰近關節分區，由於相同索引可能對應不同語意部位，難以建立一致對應，故設置為常數矩陣 $\boldsymbol{B}_h^v = 1 \in \mathbb{R}^{V' \times V'}$；對於遠距關節分區，因為位置對應固定語意，使用絕對偏差 $\boldsymbol{B}_h^v \in \mathbb{R}^{V' \times V'}$。兩者以 Kronecker 積結合為單一結構性偏差矩陣，如式(15)：

$$\boldsymbol{B}_h = \boldsymbol{B}_h^t \otimes \boldsymbol{B}_h^v \in \mathbb{R}^{T'V' \times T'V'} \tag{15}$$

最終，將 H' 個 head 的輸出結果串接，形成整體表示，如式(16)：

$$MSA(x_{msa}^{i,P}) = \text{Concat}\left(SA_1(x_{msa}^{i,P}), \cdots, SA_{H'}(x_{msa}^{i,P})\right) \tag{16}$$

在計算複雜度分析上，傳統的自注意力架構，輸入特徵張量為$x_{msa} \in \mathbb{R}^{T\times V\times C/2}$，其查詢 $(\boldsymbol{Q})$ 、鍵 $(\boldsymbol{K})$ 和值 $(\boldsymbol{V})$ 三者的維度皆為$\mathbb{R}^{(T\times V)\times C/2}$。在此情況下，標準自注意力的計算複雜度可表達如式(17)：

$$\begin{aligned} Attn &= \text{SoftMax}(\boldsymbol{QK}^{tr}) \rightarrow [(VT)^2](C/2) \\ SA(x_{msa}) &= Attn \cdot \boldsymbol{V} \rightarrow [(VT)\cdot(C/2)](VT) \end{aligned} \tag{17}$$

在第一步中，查詢與鍵張量透過點積計算出大小為 $(VT \times VT)$ 的注意力矩陣，每個元素為 $C/2$ 維的內積，因此其運算成本為 $(VT)^2 \cdot (C/2)$。在第二步中，注意力矩陣會與值張量 $\boldsymbol{V} \in \mathbb{R}^{(VT)\times(C/2)}$ 相乘，生成一個與原始值張量相同形狀的輸出，這一過程同樣需進行 $(VT)^2 \cdot (C/2)$次乘法運算。整體的總計算複雜度如式(18)：

$$2(VT)^2\frac{C}{2} \tag{18}$$

為降低自注意力的計算成本，SkateFormer 先將通道維度進行分割，並進一步結合骨架－時間分區策略，在每個子區域中獨立執行注意力運算。每組輸入特徵為 $x_{msa}^{i,P} \in \mathbb{R}^{B\times(T'\times V')\times(C/8)}$，分區後的計算僅涵蓋較小的時間與空間範圍，單一區域的注意力複雜度為 $B \cdot 2(V'T')^2 \cdot (C/8)$。四種分區的理論總計算量如式(19)：

$$\begin{aligned} P_1&: (MK)\cdot 2(LN)^2\frac{C}{8} = 2(VT)^2\frac{C}{2}\left[\frac{1}{4}\cdot\frac{1}{MK}\right] \\ P_2&: (ML)\cdot 2(KN)^2\frac{C}{8} = 2(VT)^2\frac{C}{2}\left[\frac{1}{4}\cdot\frac{1}{ML}\right] \\ P_3&: (NK)\cdot 2(LM)^2\frac{C}{8} = 2(VT)^2\frac{C}{2}\left[\frac{1}{4}\cdot\frac{1}{NK}\right] \\ P_4&: (NL)\cdot 2(KM)^2\frac{C}{8} = 2(VT)^2\frac{C}{2}\left[\frac{1}{4}\cdot\frac{1}{NL}\right] \end{aligned} \tag{19}$$

其中時間與空間維度可分別表示為 $T = M \cdot N$ 與 $V = K \cdot L$。

將四種分區的運算量加總，可得到 Skate-MSA 的總計算複雜度如式(20)：

$$2(VT)^2\frac{C}{2}\left[\frac{1}{4}\left(\frac{1}{MK}+\frac{1}{ML}+\frac{1}{NK}+\frac{1}{NL}\right)\right] \tag{20}$$

分區注意力機制 (式(20)) 相比傳統全域注意力 (式(18)) ，不僅能有效降低計算量，還能保有模型在動作語意與空間結構上的強大建模能力。

3.3.6 骨架－時間位置嵌入 Skate-Embedding (Skeletal-Temporal Positional Embedding)

為向模型傳遞結構化的時間與關節索引所對應的嵌入特徵，SkateFormer 設計一種骨架－時間位置嵌入機制 (Skate-Embedding)。如圖 24 所示，該方法結合兩種特徵：固定的時間索引特徵 (Fixed Temporal Index Features) 與可學習的骨架特徵(Learnable Skeletal Features)，分別構成時間嵌入 (Temporal Embedding, TE) 與骨架嵌入 (Skeletal Embedding, SE)，提供模型輸入所需的時空位置資訊。

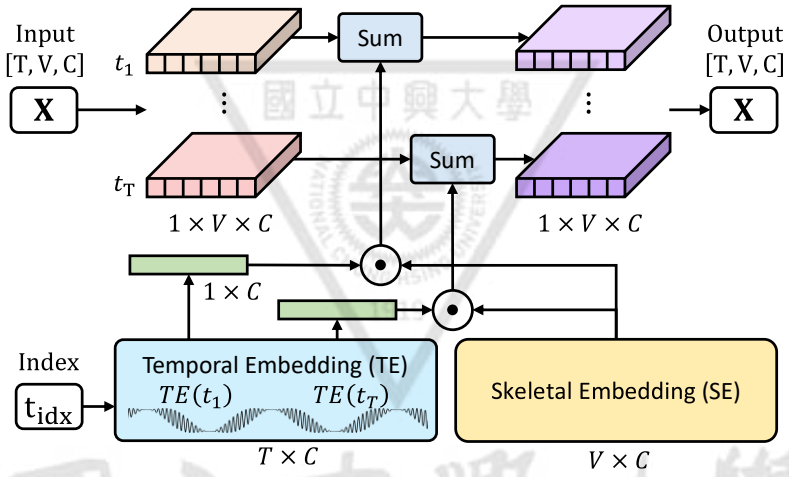


圖 24 Skate-Embedding 結構示意圖

首先，如圖 23 所示，輸入序列的時間索引可以表示為 $t_{idx} = [t_1, t_2, \cdots, t_T]$，並被正規化至 $[-1,1]$ 範圍。這些正規化後的時間索引會透過正弦曲線的位置編碼 (Sinusoidal Positional Encoding) [8] 產生固定時間嵌入 $\boldsymbol{TE} \in \mathbb{R}^{T \times C}$。另一方面，骨架嵌入並非使用實際二維座標，而是針對每個關節索引進行嵌入學習，形成可學習的關節特徵 $\boldsymbol{SE} \in \mathbb{R}^{V \times C}$，並在訓練過程中持續優化以捕捉空間語意。

最終，TE 與 SE 透過通道維度的逐元素相乘操作，構成最終的骨架－時間位置嵌入表示為 $\boldsymbol{STE} \in \mathbb{R}^{T \times V \times C}$，其定義如式(21)：

$$\boldsymbol{STE}[i,j,d] = \boldsymbol{SE}[j,d] \cdot \boldsymbol{TE}[i,d] \tag{21}$$

其中 i 表示時間索引，j 為關節索引，d 為通道維度。能使模型在每個時間點與每個關節上注入對應的語意特徵，有效保留時空結構資訊。

3.4 雙模態導向動作識別模型 (Bimodal-based Action Recognition Model)

隨著動作識別任務的應用場景日益多樣化，如何有效融合雙模態資訊以提升模型對微小動作差異的判別能力，已成為近年來的研究焦點。尤其在如桌球等細粒度動作識別任務中，動作變化往往存在於肢體微幅差異與操作物體的細節變動，傳統僅仰賴單一模態 (如 RGB 或 Skeleton) 之辨識架構，難以全面掌握時空語意脈絡與操作目標的幾何關係。

Faure 等人所提出的 Holistic Interaction Transformer Network (HIT Network) [9] 即是一具代表性的雙模態融合架構，其透過融合人物外觀影像 (RGB) 與單一幀 (single-frame) 的 2D 人體姿態特徵，並建構「人物與物件」之語意交互特徵，強調動作語意往往來自於個體對目標物的操作與操控。

然而，HIT Network 採用的骨架模態僅來自於單一幀，缺乏跨時間的動作演化建模能力，使其在處理連續性動作或細微變化的時序關係時仍有所侷限。該架構僅依賴單一時刻的骨架姿態進行語意建構，忽略了關節運動在時間上的演進模式與動作階段之間的過渡關係。由於該方法無法捕捉關節在時間軸上的動態變化，其建構的語意特徵多半僅反映當下瞬時姿態，而未能整合前後幀之間的動作趨勢與動作上下文，導致模型在辨識高相似性、具連續性特徵的動作時較易混淆，限制其在細粒度動作分類任務中的表現與應用範圍。

因此，本研究進一步整合骨架時空特徵與球拍模態資訊，針對擊球類型的高相似性辨識任務提出強化架構。我們延續 Faure 等人所提出之融合 RGB 與姿態資訊的網路設計概念，並進行兩項改進：其一，將原先僅使用單幀姿態資訊之作法，替換為 3.3 節中所提取的骨架時序特徵，以強化動作在時間軸上的語意連續性；其二，融合 3.1 節分割模型推導出的球拍幾何屬性，包括面積與中心座標等資訊，以輔助模型建構球拍運動軌跡與人物之間的空間關係。

整體架構如圖 25 所示，模型分別從輸入影像中擷取多個關鍵區域，包括人物 (P)、手部 (H) 與球拍 (O)，透過 RoI Align 投影至統一特徵空間後，並與球拍幾何特徵 (R)，一起進行區域級別的互動建模。各區域特徵經由卷積與嵌入模組提取視覺語意，再透過多層次的交互模組 (Person/Object/Hands/Racket Interaction) 捕捉其間語意關聯，並經由特徵選擇模組z_r、z_p過濾非重要特徵。

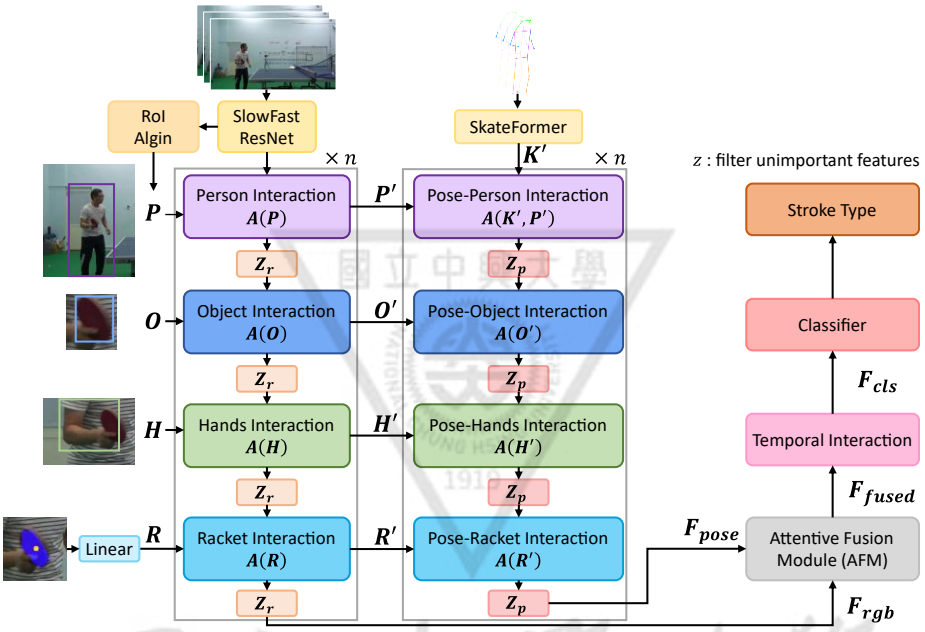


圖 25 雙模態動作識別架構圖

同時，為擷取骨架動作在時間上的連續性與空間結構，本研究將自 3.2 節所估得之多幀骨架序列，輸入至 3.3 節所採用的 SkateFormer 模型中。該模型具備時序建模能力，能有效學習人體關節點間的時空依賴關係，進而生成具語意性的骨架表徵。所得到的骨架特徵與 RGB 模態對應區域進行一對一的跨模態互動 (Pose-Person、Pose-Object、Pose-Hands、Pose-Racket)。此設計不僅強化了姿態與物件間的空間幾何對齊，也提升模型對動作語境中複雜交互關係的建模能力。

最終，模型整合骨架模態與 RGB 模態所學得之語意表徵，分別記為F_{pose}與F_{rgb}，透過注意力融合模組 (Attentive Fusion Module, AFM) 生成跨模態融合特徵F_{fused}。該融合表徵進一步輸入時序交互模組 (Temporal Interaction, TI) 以建構跨時間之語意連貫性，最終產生分類向量 F_{cls}，並輸入分類器進行動作類型的預測。

3.4.1 RGB 影像特徵擷取 (Feature Extraction from RGB Images)

本研究於 RGB 影像模態中採用 SlowFast 網路架構作為特徵提取的骨幹網路，以強化對影片中不同行為時序特性的感知能力。SlowFast 架構最初由 Feichtenhofer 等人提出[21]，其核心設計理念在於同時處理影像序列中的長期與短期時間資訊，進而提升對動作語意與細節變化的建模能力。整體架構如圖 26 所示。

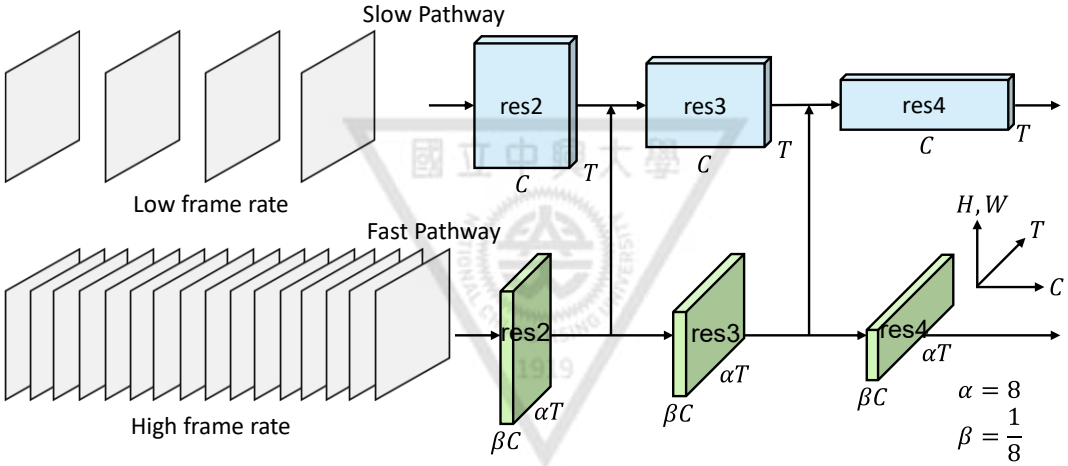


圖 26 SlowFast ResNet 架構示意圖

SlowFast 模型由兩條路徑構成：Slow Path 與 Fast Path。Slow Path 為主幹網路，負責處理低影格率 (low frame rate) 之輸入影像序列，具備較高的通道容量與較低的時間解析度，專注於擷取動作的長期時間特徵與語意脈絡；相對地，Fast Path 則以高影格率 (high frame rate) 輸入為特徵，保留豐富的時間細節變化，並透過輕量化設計 (如較少通道數) 快速捕捉短期運動變化與瞬間動作資訊。

在實作上，Fast Path 的通道數僅為 Slow Path 的 1/8 ($\beta = 1/8$)，但其時間維度則為 Slow Path 的 8 倍 ($\alpha = 8$)，實現時間解析度與計算效率之間的平衡。兩分支皆採用共享設計原則，並以 ResNet 為基礎骨幹進行多層特徵抽取 (res2、res3、res4)，在這些階段，Fast Path 所擷取之時間細節將透過 lateral connection 注入至 Slow Path，使兩條路徑得以進行跨時間尺度的資訊融合，整合長期語意與短期變化之特徵，以強化模型對複雜動作時序的表徵與辨識能力。

3.4.2 感興趣區域對齊 (Region of Interest Alignment)

本研究採用 He 等人於 Mask R-CNN 中所提出之感興趣區域對齊 (RoI Align) 技術 [12]，可以精準擷取個別對象在特徵圖上的表徵資訊。如圖 27 所示，相較於傳統 RoI Pooling 所造成的空間對齊誤差，RoI Align 移除量化操作，改以浮點數精度對位置進行對齊，有效提升特徵抽取的準確性，特別適用於精細結構辨識任務如人體姿態估計。

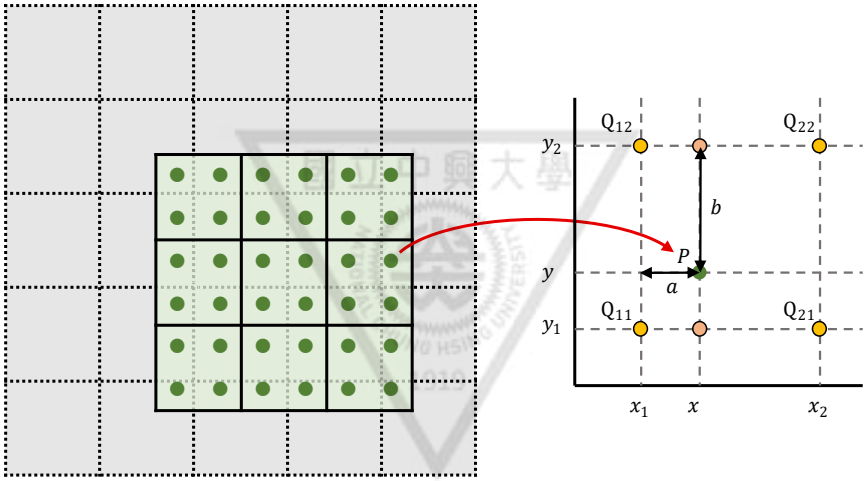


圖 27 感興趣區域對齊示意圖

RoI Align 採用雙線性插值法 (Bilinear Interpolation)，於每個 RoI bin 中對浮點座標進行採樣，計算其對應的像素值，如式(22)：

$$(x, y) = (1-a)bQ_{11} + (1-a)(1-b)Q_{12} + abQ_{21} + a(1-b)Q_{22} \quad (22)$$

其中，Q_{11}、Q_{12}、Q_{21}、Q_{22} 為 P 鄰近的整數座標點，$a = x - x_1$，$b = y - y_2$，代表浮點位置與其左上角格點的相對距離。透過上述插值操作，RoI Align 能保留精細的空間資訊，有效提升後續特徵學習與辨識的表現。

此外，本研究於每個影格中皆設置對應的動態 RoI (Dynamic RoI)，涵蓋包含人物 (Person)、手部 (Hands)、球拍 (Racket) 等不同實體區域，並對其特徵進行個別抽取。此機制能確保每一畫面中重要物件之資訊皆被完整保留，進而提升動作分析與跨模態建模的精確性，同時也為後續 RGB 與姿態模態中的語意交互與跨模態對齊操作 ($A(P)$、$A(O)$、$A(H)$、$A(R)$等模組) 提供一致且對齊良好的特徵基礎。

3.4.3 交互模組 (Interaction Module)

在動作辨識任務中，人物與其所接觸的語意區域 (如物件、手部與球拍) 之間的互動關係，常蘊含關鍵性資訊。為有效捕捉這些互動特徵，模型於 RGB 模態中構建多組語意交互模組 (Interaction Modules)，針對選定對象進行語意層級的特徵對齊與關聯建模。

此外，為強化 HIT Network 於球拍相關動作的識別能力，本研究進一步設計「球拍幾何交互模組 (Racket Interaction Module)」，作為針對擊球任務所提出的功能擴展。

整體交互模組涵蓋：人物交互 (Person Interaction) 模組，用於建構人物間的語意關係；物件交互 (Object Interaction) 模組，處理人物與周圍物件的互動語境；手部交互 (Hands Interaction) 模組，專注於人物手部與其手持物件間的精細操作資訊；以及本研究引入的球拍幾何交互 (Racket Interaction) 模組，用以整合球拍於不同時間點之幾何資訊 (如中心位置與面積)，提供補充的動作辨識依據。各模組皆基於交叉注意力 (Cross-attention) 機制進行設計，並以人物區域特徵作為查詢 (Query)，其餘語意區域則分別作為鍵 (Key) 與值 (Value) 進行配對與資訊交換，如圖 28 所示。

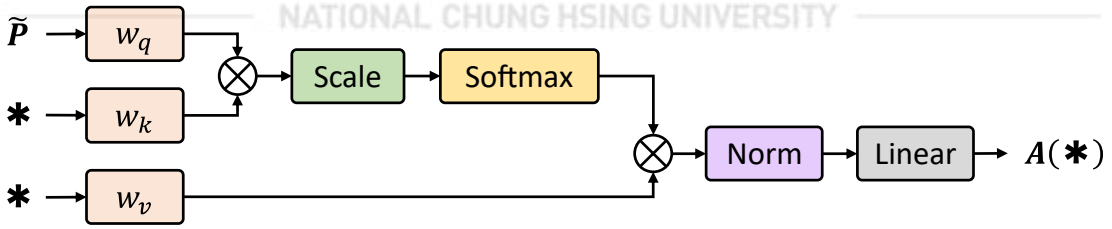


圖 28 語意交互模組的運算機制圖

整體運算流程如式(23)所示，模型依序對人物$A(P)$、物件$A(O)$、手部$A(H)$與球拍$A(R)$進行語意交互，每一階段輸出均更新為新的中間特徵 z_r，最終形成整體的 RGB 語意特徵：

$$F_{rgb} = (A(P) \Rightarrow z_r \Rightarrow A(O) \Rightarrow z_r \Rightarrow A(H) \Rightarrow z_r \Rightarrow A(R) \Rightarrow z_r) \qquad (23)$$

其中，每一模組 $A(*)$ 中的交叉注意力運算，皆透過如下之注意力計算式，如式(24)所示，人物特徵 $\tilde{P}$ 作為查詢向量，並與各模塊輸入特徵之鍵 (Key) 與值 (Value) 進行權重加權：

$$A(*) = \text{SoftMax}(\frac{w_q(\tilde{P}) \times w_k(*)}{\sqrt{d_r}}) \times w_v(*) \tag{24}$$

其中，w_q、w_k、w_v 分別為查詢、鍵與值的線性投影矩陣，d_r 為 RGB 特徵維度，透過此設計可有效聚焦於高相關性的語意區域，強化模型對區域互動資訊的感知能力。

為進一步強化語意特徵的辨識能力，濾除低貢獻度的冗餘資訊，各交互模組所輸出的特徵將傳遞至模態內部聚合模組 (Intra-modality Aggregation Module)，透過加權機制動態整合人物本體與各語意區域的交互結果，進一步突顯與動作判斷高度相關的語意資訊，如圖 29 所示。

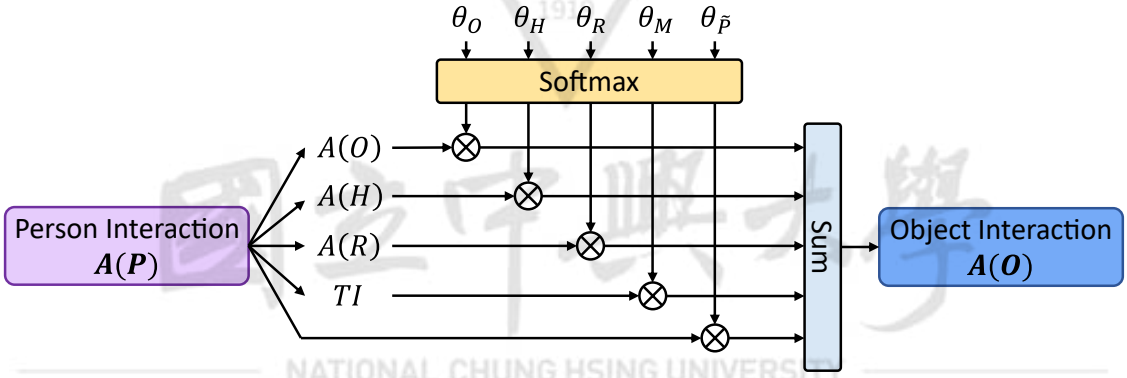


圖 29 模態內部聚合模組運算機制圖

該模組引入一組可學習的權重參數θ，根據人物本體與各語意區域在動作判斷中的貢獻程度，動態調整其交互特徵的加權比例，進一步引導模型聚焦於語意價值較高的區域資訊。其運算流程如式(25)：

$$z_r = \sum_b A(b) \times \text{SoftMax}(\theta_b), \qquad b \in (\tilde{P}, O, H, R, M) \tag{25}$$

其中，$\tilde{P}$ 為聚合後之人物特徵，M 為相鄰畫面的記憶特徵。此加權策略使模型能夠自動聚焦於高語意價值的特徵表示，並且有效提升整體的感知穩定性與識別準確度。

在姿態模態中，延續前述之語意交互設計，並以 3.3 節所輸出的骨架時序特徵 K' 為輸入，依序與 RGB 模態中對應的語意區域特徵：人物(P')、物件(O')、手部(H')與球拍(R')進行跨模態語意交互。整體運算流程如式(26)所示，並且各模組透過交叉注意力機制實現語意對齊，其計算方式如式(27)所示：

$$F_{pose} = (A(K', P') \Rightarrow z_p \Rightarrow A(O') \Rightarrow z_p \Rightarrow A(H') \Rightarrow z_p \Rightarrow A(R') \Rightarrow z_p) \quad (26)$$

$$A(K', P') = \text{SoftMax}(\frac{w_q(K') \times w_k(P')}{\sqrt{d_p}}) \times w_v(P') \quad (27)$$

其中，K'為骨架時序特徵，P'、O'、H'、R'為 RGB 模態中$A(P)$、$A(O)$、$A(H)$及$A(R)$的輸出結果，d_p為姿態模態之特徵維度，z_p為姿態模態內的聚合模塊。

3.4.4 注意力特徵融合模組 (Attentive Feature Fusion Module)

在雙模態動作識別任務中，不同模態間的訊息來源往往具有互補性。若能有效結合，便有機會建立更具判別性的語意表徵，進而提升模型對複雜動作的理解能力。然由於兩模態在語意空間與表徵形式上存在顯著差異，單純的元素相加 (Element-wise Addition) 或串接 (Concatenate) 方法難以捕捉其間的語意對齊與交互關聯。

為此，HIT Network 使用一組注意力特徵融合模組 (Attentive Feature Fusion Module, AFM) 來整合姿態分支所學得之動作表徵F_{pose}與 RGB 分支所輸出之外觀特徵 F_{rgb}。模組結構如圖 30 所示。

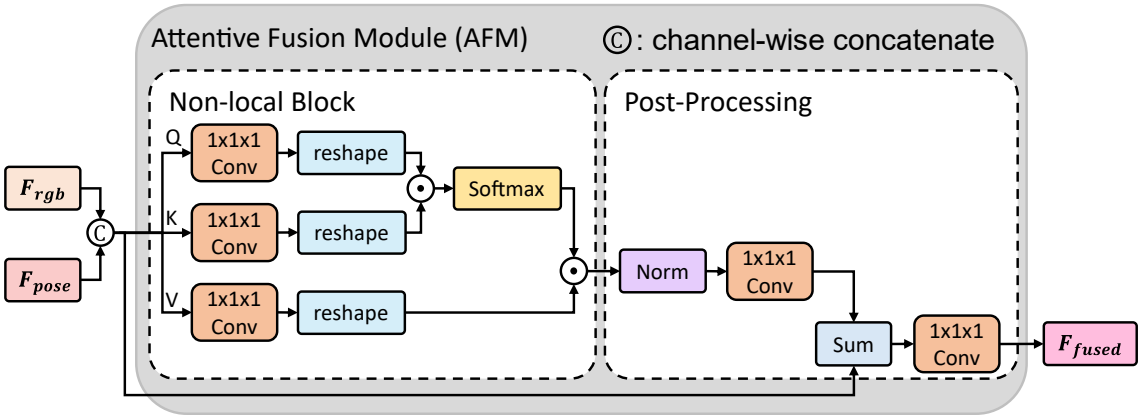


圖 30 注意力特徵融合模組之結構圖

融合流程包含兩階段：模態對齊與語意細化。首先，AFM 採用通道串接將F_{pose}與F_{rgb}結合，使原始特徵在同一空間中共存，保留各自模態特有之訊息。接著，為提升模態融合品質，模組引入自注意力機制 (Self-Attention) 以自動學習模態間的依賴關係與關鍵區域對應性。具體而言，串接後的特徵將透過三組 $1\times1\times1$ 卷積層映射為查詢 (Query)、鍵 (Key) 與值 (Value) 向量，經由 reshape、內積與 softmax 運算計算注意力權重，進行語意加權後形成初步融合特徵。上述過程可形式化表示如式(28)及式(29)：

$$F_{fused} = \Theta_{fused}(Self\ Attention(F_{pose}, F_{rgb})) \tag{28}$$

$$Self\ Attention = \text{SoftMax}(\frac{w_q \cdot w_k}{\sqrt{d_{fused}}}) \cdot w_v \tag{29}$$

其中，w_q、w_k、w_v為可學習的線性投影矩陣，負責將輸入特徵對應至注意力空間，d_{fused}為融合後特徵的通道維度。注意力運算完成後，特徵將進入後處理模組，包含 Layer Normalization、殘差連接與兩層 $1\times1\times1$ 卷積層，使融合後的語意資訊穩定整合並具備更強的區辨能力。

最終輸出之融合表徵 F_{fused}將作為後續時序建模單元的輸入基礎，協助模型捕捉長短期動作變化，並強化對微動作與高相似姿態的辨識能力。

3.4.5 時序交互模組 (Temporal Interaction Module)

我們得到了具有當前畫面之富含空間語意的特徵圖F_{fused}後，尚不足以完整描述整段動作的語意全貌。由於人類動作具有連續性，其語意常依賴跨時間的上下文資訊，僅觀察單一畫面容易導致語意斷裂與判斷偏誤。因此，長期時間特徵對動作理解至關重要，不僅有助於消除瞬時特徵的不確定性，亦能補足因遮擋、模糊所造成的資訊缺失，並協助模型掌握動作演化的整體脈絡與發展趨勢。透過整合長時間範圍內的語意線索，模型能更全面地理解動作的語境，進而提升辨識的準確性與穩定性。

為捕捉動作演化過程中的時間依賴性，HIT Network 進一步引入時序交互模組 (Temporal Interaction Module, TI)，以建構具跨時間語意連貫性的表徵。該模組整合當前畫面與其前後多幀中的人物語意資訊，強化模型對動作脈絡的理解能力與分類穩定性。

TI 模組的輸入為一段長度為 $2S+1$ 的時序融合特徵序列 M，包含過去與未來多幀畫面中目標人物所對應的融合特徵 F_{fused}。該模組採用交叉注意力機制，將當前時間步之融合特徵 F_{fused} 作為查詢向量 (Query)，並將整段記憶序列 M 作為鍵 (Key) 與值 (Value) 進行語意加權與資訊聚合。其運算流程可形式化定義如式(30)：

$$F_{cls} = TI\left(F_{fused}, M\right) \tag{30}$$

其中，F_{cls} 為最終時序語意表徵，後續將輸入至分類器進行動作辨識；M 則為跨時間的記憶特徵序列。在本實驗設定中，時間視窗參數 S 設定為 15，亦即該模組將整合來自前後共 31 幀畫面之語意資訊。

分類階段採用兩層全連接層 (Fully Connected Layer) 組成之分類器，並於各層間引入整流線性單位 (Rectified Linear Unit, ReLU) 作為非線性激勵函數，用以提升模型的表達能力與決策邊界的靈活度。分類器將時序交互模組所輸出的特徵向量 F_{cls} 作為輸入，最終輸出對應的動作類別預測結果 y_{pred}。上述映射過程與操作可分別表示如式(31)和式(32)：

$$y_{pred} = g(F_{cls}) \tag{31}$$

$$f(x) = \max(0, x) \tag{32}$$

其中，$g(\cdot)$ 表示分類器映射函數，$f(x)$ 則為 ReLU 函數，於正值區間保留訊號，在負值區間進行抑制，從而提升整體模型的非線性表達能力與訓練穩定性。

第四章 實驗結果與討論

本章將依序說明本研究的實驗設置與評估結果。4.1 節介紹實驗環境與硬體設備，4.2 節說明所使用之資料集與標註方式，0 節說明評估指標，4.4 節比較本研究與既有方法之辨識表現，並詳述各項實驗結果與分析，最後第 4.5 節呈現本研究所生成之視覺化結果，協助驗證模型在實際應用中的可行性與解釋性。

4.1 實驗環境 (Experimental Environments)

本研究利用深度學習模型進行桌球擊球動作之訓練與預測，流程涉及大量影像處理與模型計算，故採用圖形處理器 (GPU) 加速訓練與推論運算。實驗環境為 Ubuntu 22.04.1，搭配 Intel® Core™ i9-9820X 處理器、NVIDIA TITAN RTX 圖形處理器與 128GB 記憶體，以支援模型特徵計算。開發平台為 Anaconda ，採用 Python 3.10.13、PyTorch 2.4.1 與 CUDA 12.1。影像資料由 Sony HDR-CX450 高解析度數位攝影機擷取，具備每秒 60 幀 (FPS) 拍攝能力與 1920×1080 的解析度，能清楚捕捉桌球擊球過程中的細節動作。實驗環境如表 1 所示。

表 1 實驗環境設備及版本

設備/系統	版本
處理器 CPU	Intel® Core™ i9-9820X CPU @ 3.30GHz
圖形處理器 GPU	NVIDIA TITAN RTX (TU102 架構)
記憶體 Memory	128GB
作業系統 Operating System	Ubuntu 22.04.1
CUDA	12.1
Python 套件	Python 3.10.13、Pytorch 2.4.1
攝影機 Camera	Sony HDR-CX450

4.2 資料集 (Datasets)

本研究採用兩組資料集：JHMDB 用以驗證模型於通用動作辨識任務中的表現，另一組桌球擊球資料集則評估其辨識實際運動中細微動作差異的能力。

4.2.1 JHMDB

為驗證本研究方法於動作辨識任務中的準確性與時序感知能力，並參照 HIT Network [9]的實驗設置，我們採用 JHMDB (Joint-annotated Human Motion Data Base) 資料集[22]作為評估基準。如圖 31 所示，JHMDB 為動作辨識常用資料集，涵蓋 21 類動作，共 960 段裁切與標準化處理的短影片，總計 31838 幀皆具人工標註，影像解析度為 320×240，且每段影片僅呈現單一動作。本研究採用官方提供之 split-1 分割設定，並以幀級平均精度 (frame-level mAP) 作為主要評估指標，以確保實驗條件與 HIT Network 一致，並檢驗模型對時序語意與動作辨識的能力。

圖 31 JHMDB 之範例影格

本研究使用之人物、物件邊界框與人體骨架資料，皆採自 HIT Network [9] 論文所提供的偵測結果。針對人物偵測，HIT Network 採用 Köpüklü 等人 [23] 所提出之 YOWO 方法，對資料集中所有關鍵影格 (keyframes) 進行偵測，產出人物邊界框。物件偵測則使用 Faster R-CNN [24] 模型，其主幹網路為 ResNet-50-FPN [18, 25]，並於 ImageNet [26] 預訓練後，在 MSCOCO [27] 資料集上進行微調。

在人體骨架資料方面，HIT Network 採用 Detectron [28] 平台中的姿態模型，該模型以 ImageNet [26] 預訓練之 ResNet-50-FPN [18, 25] 作為主幹網路，並結合 Faster R-CNN 架構所產生的預先生成區域候選框 (Region Proposal Network, RPN) [24]， 於 MSCOCO keypoints 資料集上進行微調，以強化對人體關節位置的預測能力。該模型可針對每張關鍵影格中的每位偵測到人物，輸出 17 個符合 COCO 格式的關節點座標。

為提升姿態與人物邊界框之間的一致性與準確性，HIT Network 在訓練階段將關節座標對齊至畫面中人工標註的人物邊界框，測試階段則對齊至 YOWO [23] 所偵測出的人物邊界框。此外，於手部區域建構方面，該方法擷取左右手腕兩個關節點並建立邊界框，以強調手部區域及其間的動作，進一步增強人物與物件之間的語意建模能力。

4.2.2 桌球擊球資料集 (Table Tennis Stroke Dataset)

本研究所使用之桌球擊球資料集由國立中興大學運動與健康管理研究所許銘華教授提供，內容為教授親自示範，涵蓋正手與反手共八類典型擊球動作，如表 2 所示，包括反手切球、反手擰球、反手推球、反手拉球、正手切球、正手平擊、正手殺球以及正手拉球，並以 Sony HDR CX450 高解析度攝影機自面對球桌長邊一側錄製左側選手 (右手持拍) 的擊球過程。

表 2 資料集中所含的桌球擊球動作

反手	正手
反手切球	正手切球
反手擰球	正手平擊
反手推球	正手殺球
反手拉球	正手拉球

為真實呈現比賽情境中針對不同來球所產生的動作差異，本研究依據擊球類型調整發球機參數進行資料蒐集。實驗採用奧奇牌 TW-2700-E7K [29] 發球機，具備旋轉類型、速度、頻率、落點與角度等多項調節功能。我們針對反手推球、反手拉球、正手平擊、正手殺球與正手拉球使用設定 1；反手切球與正手切球採用設定 2；反手擰球則使用設定 3。為貼近實戰節奏，出球頻率設定為每分鐘 40 球，不同擊球動作所對應的發球機參數如表 3 所示。

表 3 不同出球類型對應之發球機參數設定

出球設定	頻率	上旋參數	下旋參數	落點區域
設定 1	40	6.0	0.8	5
設定 2	40	0.5	6.2	5
設定 3	40	0.5	5.5	5

進一步地，我們將錄製的影像採用滑動視窗 (sliding window) 策略，將每段擊球影像序列切割為固定長度的視窗片段。如圖 32 所示，我們將視窗長度設為對應於時間點前後各 16 幀 (frames)，可使模型在預測當前時間點的同時，同步考量動作的過去與未來語境。此設計有助於模型捕捉擊球動作的變化，強化對連續動作的辨識與分類能力。

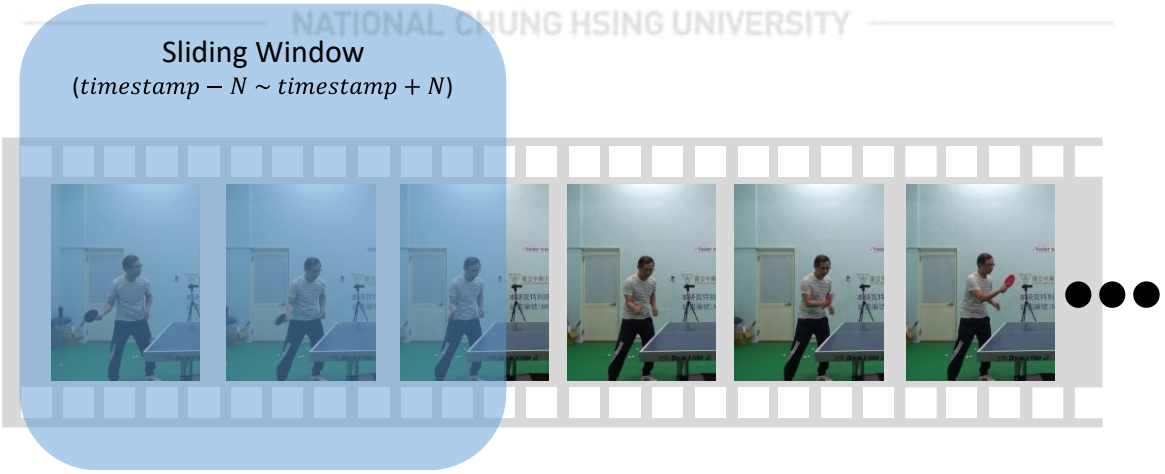


圖 32 滑動視窗切割示意圖

本研究僅針對每次擊球中語意最鮮明的前進階段 (forward phase) 進行分類標註，分為八種擊球類別，作為動作辨識的主要區間。至於收拍階段 (return phase)，因其語意相對不明顯，則統一標註為背景 (background)。

整體資料集經滑動視窗處理後，前進階段共產生 8503 筆標註樣本，其中 6250 筆用於訓練，2253 筆用於測試。收拍階段在訓練階段中，從各擊球類別中平均擷取與前進階段相同數量的樣本，共計 6250 筆；測試階段則保留收拍階段的完整序列，產生 3920 筆樣本，以真實評估模型對非動作階段的辨識能力。

表 4 資料集之各類別分布

擊球動作	**數量**
反手切球	1023
反手擰球	1002
反手推球	1604
反手拉球	829
正手切球	1049
正手平擊	609
正手殺球	901
正手拉球	1486
總計	**8503**

表 5 資料集中的資料分布

資料集	**數量**
訓練資料 (前進階段)	6250
訓練資料 (收拍階段)	6250
測試資料 (前進階段)	2253
測試資料 (收拍階段)	3920

本研究使用第 3.1 節所提之球拍實例分割結果，於每幀影像中擷取球拍遮罩的面積、中心座標與物件框，作為幾何特徵輸入模型，進一步輔助辨識揮拍動作的發起時機，如圖 33 所示。面積可反映球拍在視野中的出現程度，而中心座標則標示其於空間中的位置，兩者皆有助於提供額外的時序線索，強化模型對動作起始階段之感知能力，提升動作邊界處的分類準確性。

(a) (b) (c) (d)

(e) (f) (g) (h)

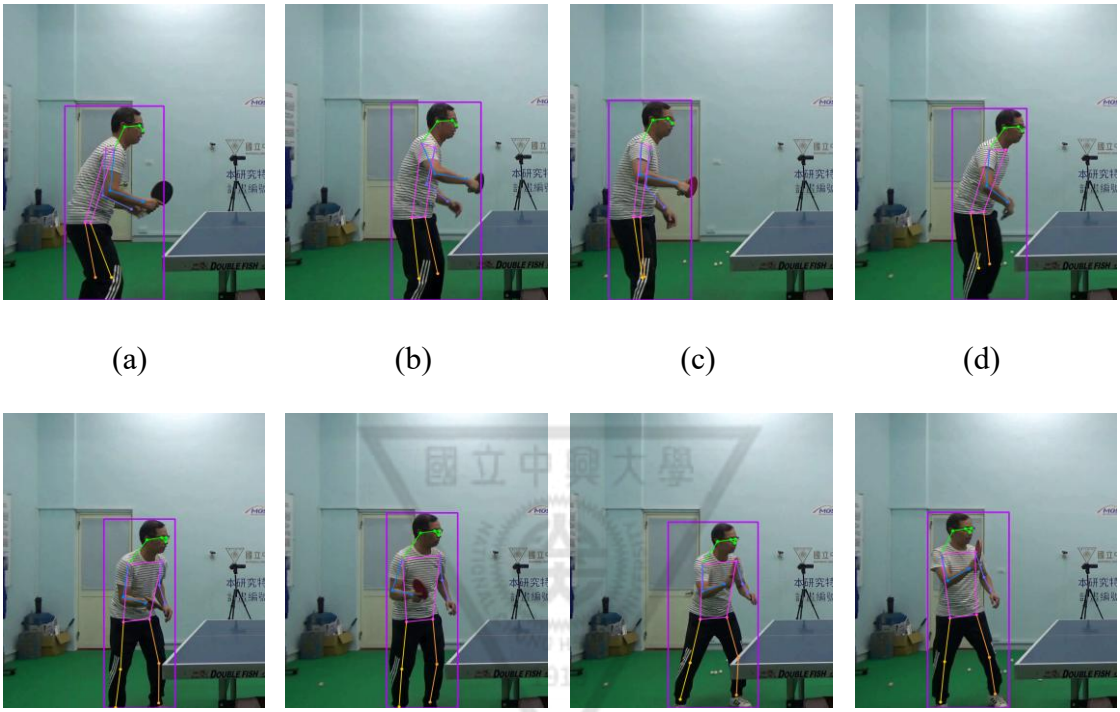
圖 33 球拍遮罩面積、中心位置與物件框之視覺化結果

(a) 反手切球 (b) 反手擰球 (c) 反手推球 (d) 反手拉球

(e) 正手切球 (f) 正手平擊 (g) 正手殺球 (h) 正手拉球

本研究採用 Ultralytics [14] 官方釋出之 YOLOv11-Pose 模型，執行擊球影像中的 2D 人體姿態估計。該模型具備同時偵測人框與骨架關節點的能力，並直接於本研究自建之擊球資料集上進行推論，以取得每幀影像中選手的骨架關節點位置。透過此估計結果，模型得以掌握肢體動作變化，進一步擷取動作演化的時序特徵，強化對不同擊球類型的辨識能力。

如圖 34 所示，YOLOv11-Pose 可穩定預測包含頭部、軀幹與四肢等關節點，提供後續骨架模態建模之基礎資訊。



(a) (b) (c) (d)

(e) (f) (g) (h)

圖 34 2D 骨架估計結果

(a) 反手切球 (b) 反手擰球 (c) 反手推球 (d) 反手拉球

(e) 正手切球 (f) 正手平擊 (g) 正手殺球 (h) 正手拉球

4.3 評估標準 (Evaluation Metrics)

本研究採用混淆矩陣 (Confusion Matrix) 作為主要評估依據，用以衡量動作分類模型在多類別辨識任務中的預測表現。如圖 35 所示，混淆矩陣由四個基本構成元素組成：真陽性 (True Positive, TP)、偽陽性 (False Positive, FP)、偽陰性 (False Negative, FN) 與真陰性 (True Negative, TN)。其中，TP 表示模型正確辨識出屬於特定類別的樣本；FP 表示模型誤將非該類別樣本預測為該類別；FN 則為實際為該類別但模型未能正確預測者；而 TN 則代表模型正確排除非該類別樣本。

透過混淆矩陣，可進一步觀察模型在各類別間的分類誤差，掌握易混淆的動作組合。此評估方式能有助於深入分析模型在特定擊球類別上的辨識盲點，特別是在多類別任務中，混淆矩陣提供更細緻的診斷依據，作為模型優化的重要參考。透過矩陣中不同類別之間的誤判關係，研究者可更清楚地掌握模型對各類別的學習情況，進而調整訓練策略與增強特定類別特徵，以提升整體辨識性能與類別間的區辨能力。混淆矩陣所呈現的視覺化結果，也能作為後續定性分析的重要輔助工具。

Confusion Matrix		Prediction	
		Positive	Negative
Ground Truth	Positive	True Positives (TP)	False Negatives (FN)
	Negative	False Positives (FP)	True Negatives (TN)

圖 35 混淆矩陣

另外，為全面評估模型於多類別擊球辨識任務中的整體表現，本研究採用三項常見分類指標：精確率 (Precision)、召回率 (Recall) 與 F1-score，分別對應模型的預測正確性、檢出能力與整體平衡性。

Precision (精確率) 衡量模型預測為某類別的樣本中，有多少實際屬於該類別，能反映模型產生誤判的情形，特別是在誤將其他類別誤分類為目標類別時。其公式如式(33)：

$$Precision = \frac{TP}{TP + FP} \tag{33}$$

Recall (召回率) 則表示在所有實際屬於某類別的樣本中，有多少被模型成功預測出來，評估模型的檢出能力。若 Recall 偏低，可能代表模型遺漏了大量應該被辨識的樣本。其計算方式如式(34)：

$$Recall = \frac{TP}{TP + FN} \tag{34}$$

F1-score 為 Precision 與 Recall 的調和平均，用於綜合評估模型在精確性與完整性間的平衡，特別適合用於類別分布不均或需同時考量誤判與遺漏的情境。其公式如式(35)：

$$F1\text{-}score = 2 \times (\frac{Precision \times Recall}{Precision + Recall}) \quad (35)$$

上述三項指標皆以百分比形式呈現，能有效衡量模型在各類別下的預測表現，並提供分類品質的直觀數值依據。

此外，為整體量化模型的分類效能，本研究採用總和平均 (Macro Average) 作為整體指標。Macro Average 會先分別計算每一類別的 Precision、Recall 與 F1-score，然後對所有類別取平均，使所有類別在評估中具有相同權重。其公式如式(36)：

$$Macro\ Average = \frac{1}{N}\sum_{i=1}^{N} Metric_i \quad (36)$$

其中，$Metric_i$ 表示第 i 類的指標數值，N 為總類別數。此指標確保各類別權重相等，避免多數類別主導結果，有助於真實反映模型於整體任務的穩定性與平衡性。

4.4 比較結果 (Comparisons)

為驗證模型於不同場景的辨識能力，本節比較其在 JHMDB 與本研究桌球擊球資料集上的分類與時序表現。

4.4.1 JHMDB 結果分析 (JHMDB Results Analysis)

為觀察模型於時序連續影格中的預測一致性與語意理解能力，如圖 36 與圖 37 所示，分別展示 shoot ball 與 shoot bow 動作之預測序列。每個影格左上角標示模型所預測的動作類別，其中綠色文字代表預測正確，紅色文字則表示分類錯誤。

HIT Network 僅依賴單幀骨架進行辨識，缺乏跨幀語意建模能力，容易於動作過程中因關節構型相似而產生誤判。例如，在 shoot ball 動作中，起跳階段常被誤判為 jump，手臂自然下垂時被誤判為 catch，落地時因膝蓋彎曲與身體前傾則被辨識為 run；此外，shoot bow 動作中的準備與拉弓階段，亦可能被誤判為 pour 或 shoot gun。此類錯誤反映其對動作語意變化與時間連貫性的理解有限。

相較之下，本研究方法結合骨架的時序建模與影像模態的語意輔助，能穩定掌握整段動作的語意結構。從圖中可見，模型對主要動作階段皆能連續且正確地預測為 shoot ball 與 shoot bow，具有良好的一致性與時間脈絡連結，展現出優異的語意辨識與跨幀推理能力，亦於姿態轉換與動作邊界階段維持穩定預測。

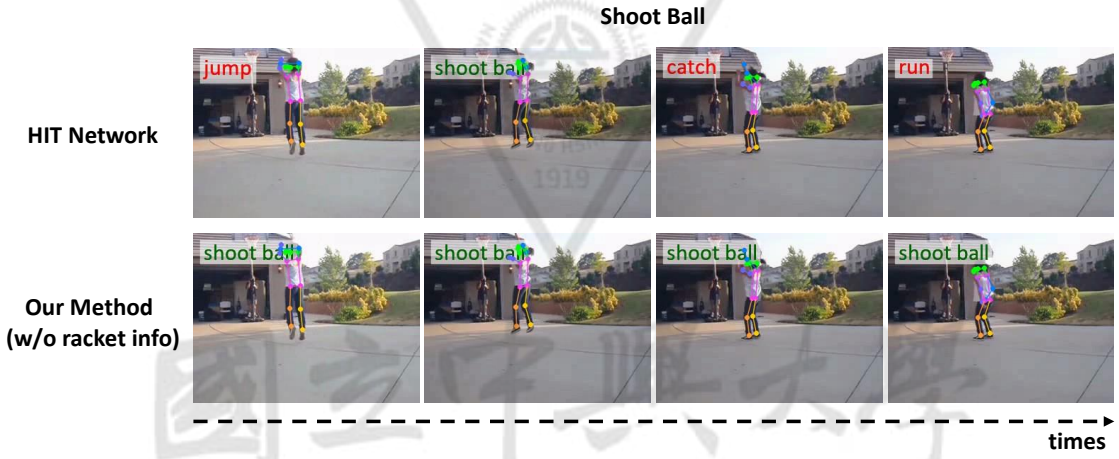


圖 36 shoot ball 動作之預測結果比較

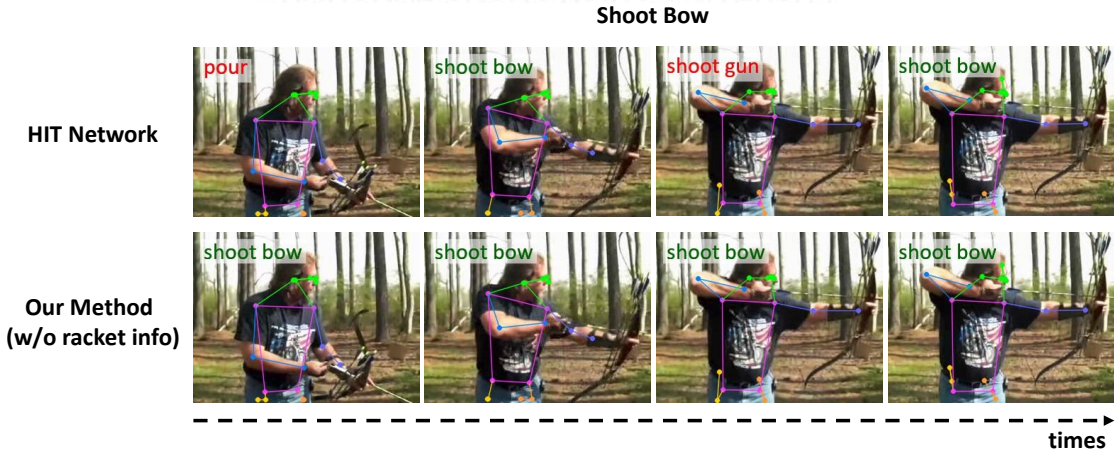


圖 37 shoot bow 動作之預測結果比較

如表 6 所示，HIT Network 在資料集上的 Precision、Recall 與 F1-score 分別為 83.9%、82.4% 與 82.4%；而本研究所提出之方法 (不含球拍資訊) 則進一步提升至 84.2%、83.8% 與 83.8%。透過多幀骨架時序建模，能有效捕捉動作語意隨時間演變的脈絡，不僅在視覺連續性上具優勢，亦於分類準確率上展現更佳表現，證實其於通用動作辨識任務中的適用性與穩定性。

表 6 資料集之分類表現比較

Model	Precision (Avg.)	Recall (Avg.)	F1-score (Avg.)
Hit Network [9]	83.9%	82.4%	82.4%
Our Proposed Method (w/o racket info)	**84.2%**	**83.8%**	**83.8%**

4.4.2 桌球擊球資料集結果分析 (Table Tennis Stroke Dataset Results Analysis)

為驗證各類動作辨識模型於本研究所建構之桌球擊球資料集上的實際分類效能，本節透過混淆矩陣、定量指標與實例分析，探討不同模型在動作語意辨識、時序建模與邊界解析等面向的表現。

分析內容依模型設計理念分為五個子節：4.4.2.1 探討 SlowFast ResNet 在 RGB 單模態下的表現與侷限；4.4.2.2 說明 HIT Network 採用單幀骨架所面臨的時序資訊不足與方向模糊問題；4.4.2.3 評估 SkateFormer 的時序建模優勢與遮蔽帶來的穩定性挑戰；4.4.2.4 說明本研究所使用跨模態架構在結構補強與分類精度上的貢獻；最後於 4.4.2.5 引入球拍幾何資訊以提升模型對動作邊界的辨識能力，進一步強化整體辨識效能與時序解析穩定性。

4.4.2.1 SlowFast ResNet 分析：單模態特徵表徵侷限

(SlowFast ResNet: Limitations of Unimodal Feature Representation)

為評估單一模態特徵對動作辨識任務之基本分類能力，本研究首先以 SlowFast ResNet 作為 RGB 影像模態之骨幹網路進行實驗。SlowFast 架構透過雙路徑策略，兼顧動作之空間與時間特徵提取，Fast branch 專注於動作細節變化，Slow branch 則保留較長時間範圍內之語意資訊。然而，該方法僅仰賴影像中所擷取之外觀與動作線索，缺乏顯式結構與動態關係建模機制，導致其在部分姿態相似之擊球動作分類上表現受限。

如圖 38 所示，SlowFast ResNet 模型於桌球擊球資料集上的混淆矩陣結果揭示其分類表現。

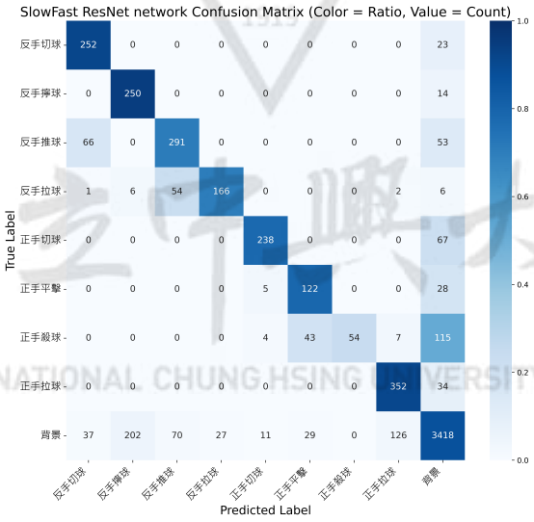


圖 38 SlowFast ResNet 模型之混淆矩陣結果

SlowFast ResNet 模型在 Precision、Recall 與 F1-score 上如表 7 所示，分別為 77.8%、76.4% 與 73.4%，雖具備一定分類能力，但表現仍有明顯進步空間。

表 7 SlowFast ResNet 模型之整體分類表現

Model	Precision (Avg.)	Recall (Avg.)	F1-score (Avg.)
SlowFast ResNet [21]	77.8%	76.4%	73.4%

4.4.2.2 HIT Network 分析：單幀骨架表示的方向模糊與時序侷限 (HIT Network: Directional Ambiguity and Temporal Limitations in Single-Frame Skeletons)

HIT Network [9] 模型為一套設計用於多模態動作辨識之網路架構，於骨架模態中僅採用單幀 (single-frame) 姿態資訊進行編碼。雖然此種設計能有效降低計算成本並保有基本的結構辨識能力，但在處理如桌球擊球這類需依賴時間變化與動作連續性的應用場景時，易面臨顯著限制。

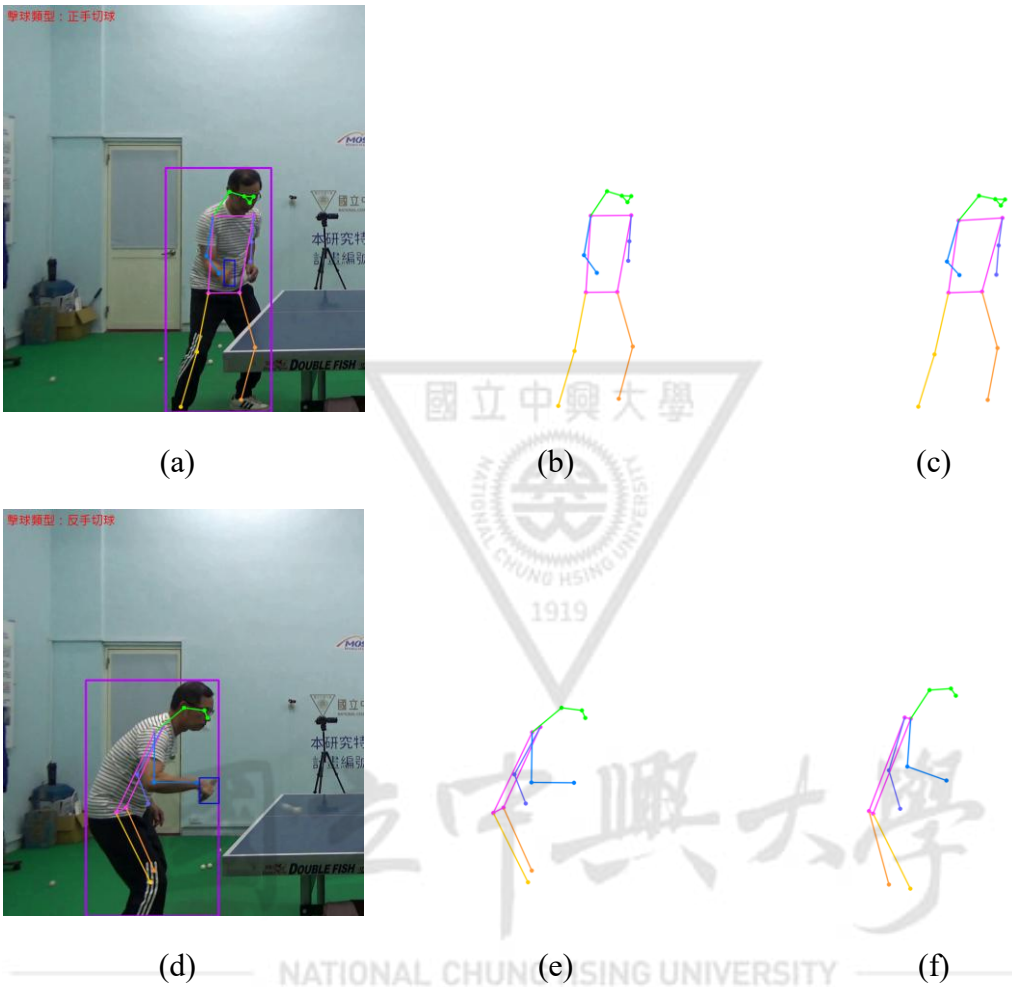
首先，單幀骨架缺乏動作方向 (motion direction) 資訊，使模型難以區分動作的前進階段與收拍階段。如圖 39 所示，(a) 為模型於反手拉球動作中之收拍階段的預測結果，可見其僅透過單張骨架姿態進行判斷，缺乏時間軸上的動作變化線索，導致動作階段無法被正確識別。比較 (b) 與 (c)，分別呈現反手拉球動作中之收拍階段的姿態與前進階段的姿態，兩者在單幀姿態上僅存在細微差異。由於模型無法掌握連續動作間的變化趨勢，造成難以推斷當前動作所處的方向與語意階段，進而產生方向性模糊 (directional ambiguity) 問題。

(a) (b) (c)

圖 39 單幀骨架姿態之方向性模糊示意圖

(a) 收拍階段之預測結果影像 (b) 收拍階段之骨架姿態 (c) 前進階段之骨架姿態

再者，如圖 40 所示，HIT Network 在骨架模態上的另一項限制為缺乏時間軸上的連續訊息，難以建立跨幀的動作脈絡 (temporal context)，當不同動作在某些關節配置上高度相似時，模型容易出現混淆。



(a) (b) (c)

(d) (e) (f)

圖 40 單幀骨架姿態之姿態歧義示意圖

(a) 測試樣本之錯誤預測結果 (GT: 正手平擊, Pred: 正手切球)

(b) 錯判樣本之骨架姿態 (正手平擊)

(c) 類似姿態之正確樣本骨架 (正手切球)

(d) 測試樣本之錯誤預測結果 (GT: 反手推球, Pred: 反手切球)

(e) 錯判樣本之骨架姿態 (反手推球)

(f) 類似姿態之正確樣本骨架 (反手切球)

例如 (a) 為其中一筆測試樣本，其實際標註為正手平擊，卻被誤判為正手切球。對應的骨架如 (b) 所示，與 (c) 中來自正手切球類別的正確樣本骨架在關節配置上幾乎一致。同樣情形也發生於 (d) 所示之測試樣本，其實際動作為反手推球，卻被誤判為反手切球，對應之骨架如 (e) 所示，亦與 (f) 中來自反手切球的正確樣本骨架極為相似。此結果顯示，單幀骨架在辨識姿態相近的細微動作時，因缺乏時序資訊而容易產生混淆。

如圖 41 所示，混淆矩陣結果亦可觀察到上述問題所導致之辨識誤差，例如反手拉球與背景之間存在錯誤分類現象，此外，正手平擊常被誤判為正手切球，反手推球亦經常被誤判為反手切球，反映模型在處理語意相近且姿態類似之動作時，缺乏有效的時序辨識機制，容易產生混淆。

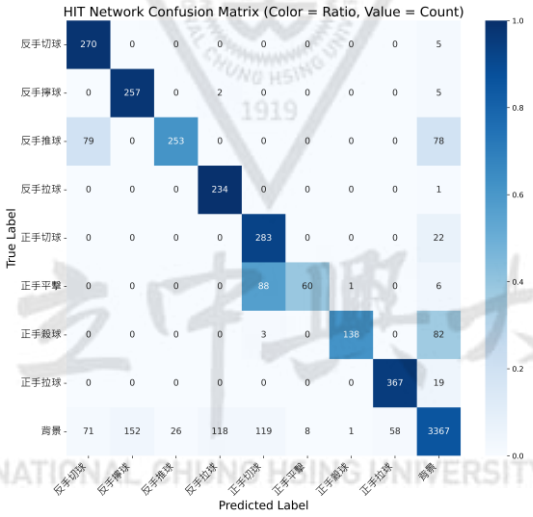


圖 41 HIT Network 模型之混淆矩陣結果

HIT Network 模型在 Precision、Recall 與 F1-score 上如表 8 所示，分別為 78.7%、81.2% 與 76.4%。雖具有一定分類能力，惟面對細節相似或需依賴時序資訊的擊球動作時，模型表現仍顯不足。

表 8 HIT Network 模型之整體分類表現

Model	Precision (Avg.)	Recall (Avg.)	F1-score (Avg.)
Hit Network [9]	78.7%	81.2%	76.4%

4.4.2.3 SkateFormer 分析：時序建模強化模型能力但關節遮蔽影響辨識穩定性（SkateFormer: Temporal Modeling Enhances Performance, but Joint Occlusion Affects Recognition Stability）

SkateFormer [7] 採用多幀骨架序列進行時序建模，專為捕捉人體動作中連續性的關節變化與時序依賴關係所設計。其核心設計包含對骨架與時間軸進行分區機制，強化局部與全域的關節動態交互，進一步提升對細緻動作變化的感知能力。此機制能有效解決傳統單幀模型在姿態相似動作下難以區分語意的問題，提升模型在複雜動作辨識任務中的表現。

如圖 42 所示，SkateFormer 相較於 HIT Network 呈現更清晰集中於對角線的混淆矩陣，顯示分類結果更加準確且具一致性。特別是在正手切球、正手平擊、反手推球等僅在肢體動作順序或揮拍幅度上存在微小差異的動作類別中，SkateFormer 藉由連續骨架序列建模有效區分其語意與動態特徵，展現優異的動作語意理解與辨識能力。

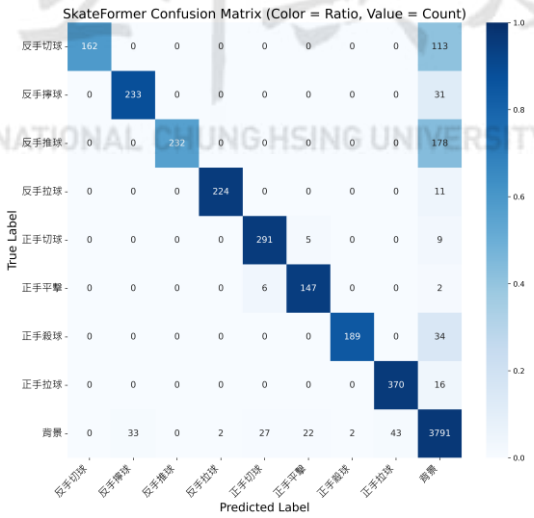
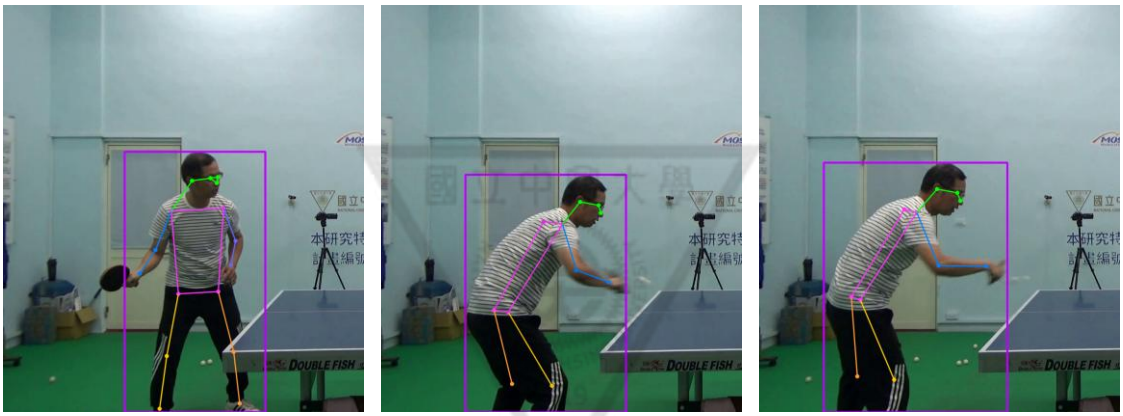


圖 42 SkateFormer 模型之混淆矩陣結果

然而，SkateFormer 亦呈現出其骨架模態的潛在侷限。由於其輸入特徵完全仰賴 2D 人體姿態估測結果，若關節因遮蔽而消失或錯位，將顯著影響模型判斷。如圖 43 所示，(a) 為正手平擊，所有關節皆穩定辨識；而 (b) 與 (c) 分別為反手切球與反手推球，皆因左手被軀幹遮蔽，導致對應節點出現辨識不穩定、偶有遺失的情形。此類遮蔽現象造成骨架結構劣化 (pose degradation)，破壞原始時序結構的連貫性，使模型難以正確解析動作語意，進而影響分類準確度。



(a) (b) (c)

圖 43 遮蔽導致關節點遺失與骨架劣化之示意圖

(a) 所有關節節點皆穩定辨識 (正手平擊)

(b) 左手因軀幹遮蔽導致節點遺失 (反手切球)

(c) 左手因軀幹遮蔽導致節點遺失 (反手推球)

整體而言，SkateFormer 模型在 Precision、Recall 與 F1-score 上如表 9 所示，分別為 93.3%、85.2% 與 87.9%，展現出優異的分類效能與時序建模能力。該模型能有效捕捉骨架間的時間依賴關係，於多數擊球類型中皆具良好區辨力，惟在遇到關節遮蔽或姿態估測誤差等實務挑戰時，仍可能對辨識穩定性造成影響。

表 9 SkateFormer 模型之整體分類表現

Model	Precision (Avg.)	Recall (Avg.)	F1-score (Avg.)
SkateFormer [7]	93.3%	85.2%	87.9%

4.4.2.4 本研究方法 (不含球拍資訊)：跨模態特徵互補提升分類效果 (Proposed Method without Racket Information: Cross-Modal Feature Complementarity Enhancing Classification)

為克服單一模態在動作辨識中所面臨的結構侷限與語意缺失問題，本研究使用一套融合骨架與影像特徵的跨模態辨識架構。骨架模態基於 SkateFormer [7] ，具備優異的時序建模能力，能捕捉關節間的時空依賴與動作結構語意；RGB 模態則採用 SlowFast ResNet [21]，能從影像中擷取肢體局部的紋理細節、輪廓變化與背景語境，補強骨架模態在關節遮蔽或估測誤差下的資訊不足。當骨架表徵品質下降時，RGB 模態可提供額外的動作判斷依據，確保模型在處理結構缺損的情境中，仍具備良好的分類穩定性與辨識準確度。

如圖 44 所示，本研究方法 (不含球拍資訊) 在混淆矩陣上展現出較集中之對角線分布，整體優於 SlowFast ResNet [21]、HIT Network [9] 與原始 SkateFormer [7] 模型，顯示分類穩定性與一致性更佳。

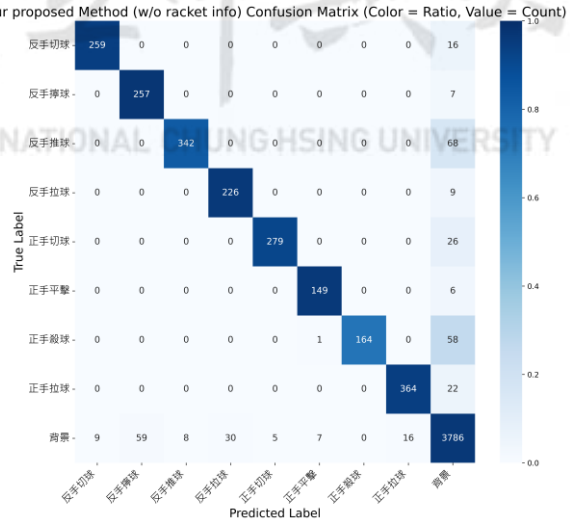


圖 44 本研究方法 (不含球拍資訊) 之混淆矩陣結果

如表 10 所示，在定量指標上，本方法於 Precision、Recall 與 F1-score 指標分別達到 94.2%、91.5% 與 92.4%，整體優於前述各模型。

表 10 本研究方法 (不含球拍資訊) 之整體分類表現

Model	Precision (Avg.)	Recall (Avg.)	F1-score (Avg.)
SlowFast ResNet [21]	77.8%	76.4%	73.4%
Hit Network [9]	78.7%	81.2%	76.4%
SkateFormer [7]	93.3%	85.2%	87.9%
Our Proposed Method (w/o racket info)	**94.2%**	**91.5%**	**92.4%**

4.4.2.5 本研究方法 (引入球拍幾何資訊)：強化動作邊界辨識 (Proposed Method with Racket Geometric Information: Enhancing Boundary-Aware Action Recognition)

在觀察擊球動作的時間分佈時，本研究發現平均每次擊球動作約涵蓋 31 張影格，如表 11 所示。

表 11 本研究方法 (不含球拍資訊) 之整體分類表現

擊球動作	平均影格數
反手切球	27.5
反手擰球	29.3
反手推球	31.5
反手拉球	33.6
正手切球	33.9
正手平擊	31.0
正手殺球	27.9
正手拉球	29.7
平均	**30.6**

為分析模型於整段時序中的辨識精度，我們將所有錯誤預測之影格依序對齊並正規化至 31 格時間軸，繪製出平均錯誤率熱圖。如圖 45 所示，模型於擊球起始與結束兩端有顯著的錯誤累積，推論模型難以準確判斷動作的邊界位置，導致預測結果出現邊界模糊 (boundary ambiguity) 之現象。

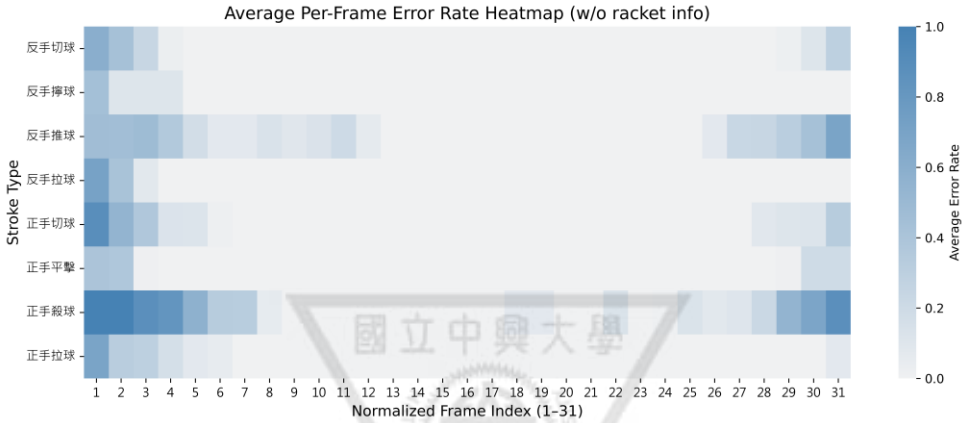


圖 45 未引入球拍資訊下模型於動作邊界處之錯誤率分布

為解決此問題，我們於模型中引入「球拍區域面積」與「中心座標」作為額外幾何特徵，輔助辨識當前影格所對應的動作階段。前進與收拍階段在球拍路徑與所佔面積上具幾何差異，僅以單幀資訊，亦可推斷動作的開始與結束時機，補強滑動視窗下缺乏明確起訖訊號的侷限，提升動作邊界處的時序解析能力與整體穩定性。

實驗結果顯示，引入球拍資訊後，模型於擊球邊界處之錯誤率明顯下降 (如圖 46 所示)，顯示幾何輔助特徵有助於準確判斷動作起始與結束時機，改善預測邊界模糊 (boundary ambiguity) 問題。

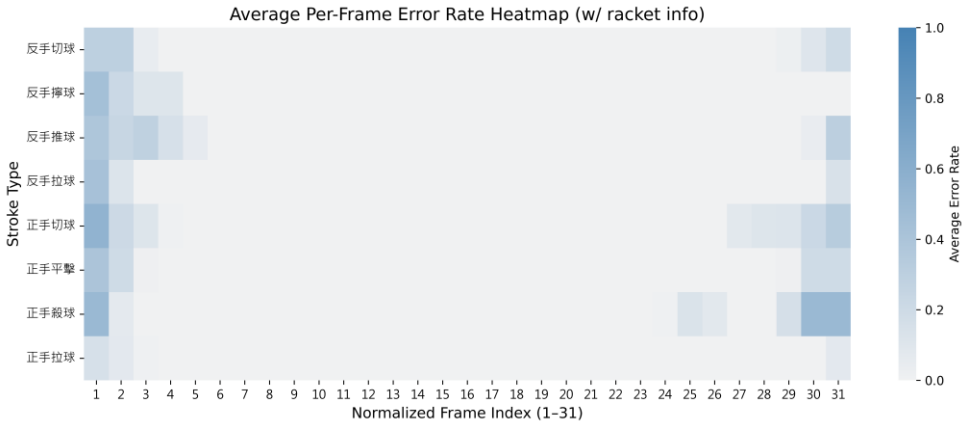


圖 46 引入球拍資訊下模型於動作邊界處之錯誤率分布

此外，整體分類效能亦有顯著提升，其 Precision、Recall 與 F1-score 如表 12 所示，分別上升至 96.1%、96.4% 與 96.2%，進一步驗證本方法於動作辨識任務中的效益。

表 12 本研究方法於引入與未引入球拍幾何特徵下之整體分類效能比較

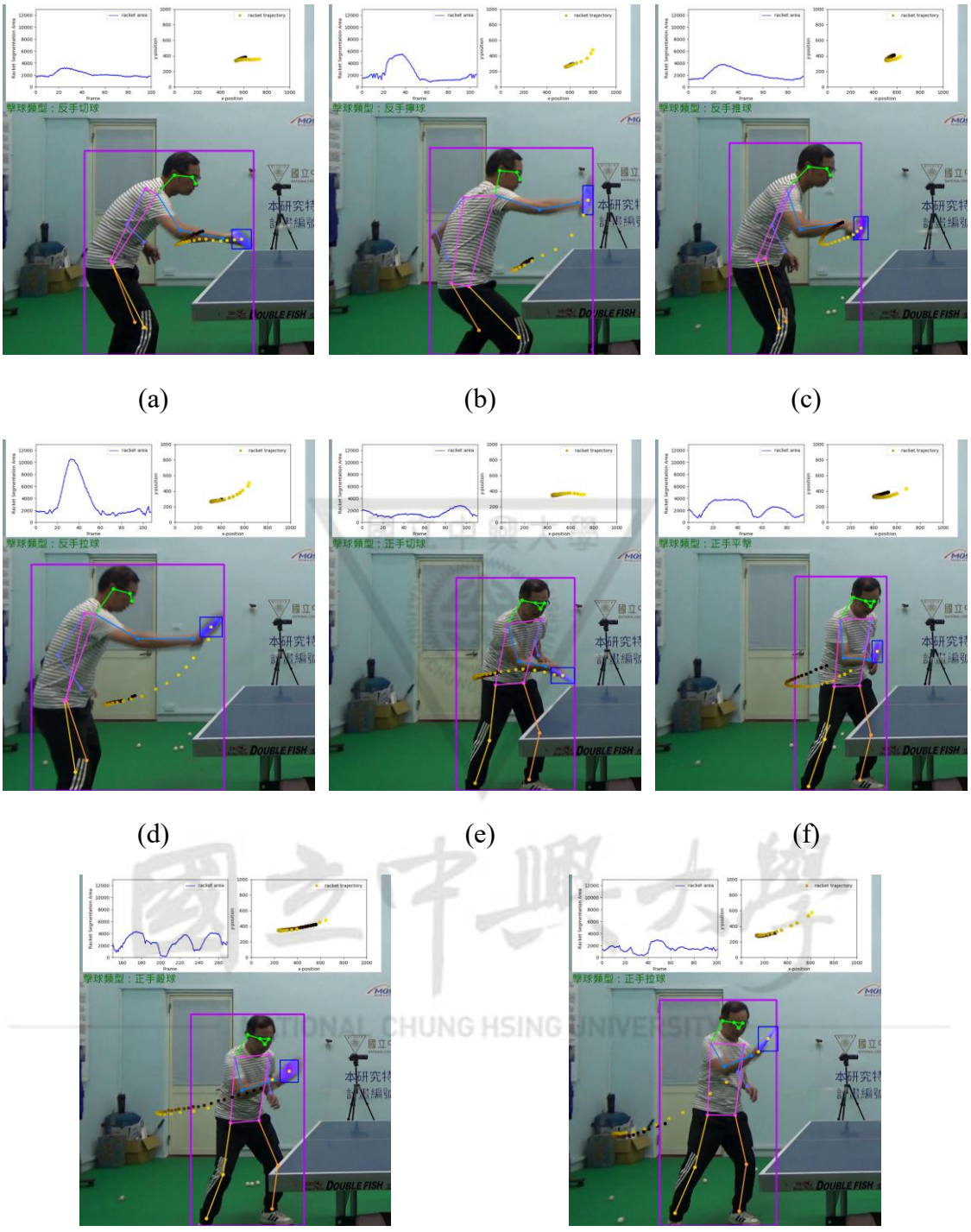
Model	Precision (Avg.)	Recall (Avg.)	F1-score (Avg.)
Our Proposed Method (w/o racket info)	94.2%	91.5%	92.4%
Our Proposed Method (w/ racket info)	**96.1%**	**96.4%**	**96.2%**

4.5 實驗結果 (Experimental Results)

為了更直觀地展示本研究之擊球動作辨識系統的實際效能，我們以視覺化方式呈現各模組之輸出結果。如圖 47 所示，我們將辨識結果疊加於影片畫面上，提供使用者辨識資訊。其中，預測出的動作類別將顯示於畫面左上角，若與標註結果相符，則以綠色文字標示，代表該影格預測正確；反之，若預測錯誤，則以紅色文字顯示，以利使用者快速判讀模型表現。

同時，為強化對動作時序與球拍幾何特徵理解，我們亦於每一影格上方繪製球拍中心位置與面積，提供模型於邊界辨識所依賴之幾何輔助資訊。此外，骨架姿態估測與球拍分割遮罩亦同步可視化，使整體預測過程與各模態輸出結果一目了然。

本研究之模型能穩定且正確地辨識正手與反手共八類典型的擊球動作，且皆能於動作進行期間輸出正確類別，顯示模型具備良好的時序辨識能力與姿勢區辨能力。此一結果驗證本研究所採用之骨架時序建模與球拍輔助幾何資訊設計，能有效強化模型對於類似動作之語意差異辨識能力，並在桌球擊球影片中展現穩定的跨類別分類能力，顯示其在智慧運動分析應用中的實用潛力與可行性。



(a) (b) (c)

(d) (e) (f)

(g) (h)

圖 47 本研究動作辨識系統之輸出結果視覺化

(a) 反手切球 (b) 反手擰球 (c) 反手推球

(d) 反手拉球 (e) 正手切球 (f) 正手平擊

(g) 正手殺球 (h) 正手拉球

第五章 結論與未來展望

本研究針對桌球擊球動作辨識任務，使用一套融合 2D 骨架與 RGB 影像之雙模態動作辨識架構。相較於過往僅依賴單幀姿態或單一模態輸入之方法，本方法結合具備時序建模能力的 SkateFormer 骨架分支與 SlowFast ResNet 影像分支，進行跨模態語意互補與特徵融合，有效提升動作語意理解與分類效能。骨架模態負責捕捉關節間時空依賴關係，而 RGB 模態則補足骨架於遮蔽或姿態估測不穩定情境下的資訊缺口，進一步強化動作辨識的穩定性與準確性。

為處理擊球動作邊界辨識困難的問題，我們進一步分析整體擊球動作的時間分佈，發現平均長度約為 31 幀，且錯誤率常集中於動作的起始與終止區段。基於此觀察，我們引入球拍區域的幾何資訊（包含面積與中心座標）作為輔助模態，有效提供動作發生的時序線索。實驗結果證實，此設計能顯著提升模型在動作邊界區間的判斷能力，使錯誤率熱圖中於時間邊界位置的誤判情形明顯減少。

在定量結果方面，所提出方法於 Precision、Recall 與 F1-score 上分別達到 96.1%、96.4% 與 96.2%，優於 SlowFast ResNet、HIT Network 與 SkateFormer 等基準模型。混淆矩陣亦顯示本方法於多數擊球類型中展現穩定且集中於對角線的預測分佈，證實融合多模態與幾何資訊確實能有效提升桌球擊球動作之辨識表現。此外，本方法亦在 JHMDB 通用動作資料集上，於 Precision、Recall 與 F1-score 上分別達到 84.2%、83.8% 與 83.8%，展現出良好的泛化能力。

展望未來，仍有數個方向可進一步優化本研究工作。首先，球拍區域目前僅使用面積與中心座標等幾何資訊進行建模，未來可延伸至更高層次的動態表徵，如揮拍速度、角度變化等時序特徵。其次，骨架模態採用 2D 姿態估測，可能受限於視角與遮蔽，未來可結合多攝影機進行 3D 姿態重建，進一步提升姿態結構的完整性與辨識精度。

參考文獻

[1] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele, "2d human pose estimation: New benchmark and state of the art analysis," in *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*, 2014, pp. 3686–3693.

[2] K. M. Kulkarni and S. Shenoy, "Table tennis stroke recognition using two-dimensional human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 4576–4584.

[3] W. Liu *et al.*, "Ssd: Single shot multibox detector," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, 2016: Springer, pp. 21–37.

[4] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5693–5703.

[5] P.-E. Martin, J. Benois-Pineau, R. Péteri, and J. Morlier, "Fine grained sport action recognition with Twin spatio-temporal convolutional neural networks: Application to table tennis," *Multimedia Tools and Applications,* vol. 79, pp. 20429–20447, 2020.

[6] Z. Zivkovic and F. Van Der Heijden, "Efficient adaptive density estimation per image pixel for the task of background subtraction," *Pattern recognition letters,* vol. 27, no. 7, pp. 773–780, 2006.

[7] J. Do and M. Kim, "Skateformer: skeletal-temporal transformer for human action recognition," in *European Conference on Computer Vision*, 2024: Springer, pp. 401–420.

[8] A. Vaswani *et al.*, "Attention is all you need," *Advances in neural information processing systems,* vol. 30, 2017.

[9] G. J. Faure, M.-H. Chen, and S.-H. Lai, "Holistic interaction transformer network for action detection," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2023, pp. 3340–3350.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM,* vol. 60, no. 6, pp. 84–90, 2017.

[11] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[12] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[13] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "Yolact: Real-time instance segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9157–9166.

[14] Ultralytics, "Ultralytics YOLOv11," 2024. [Online]. Available: https://github.com/ultralytics/ultralytics.

[15] G. Jocher, "Ultralytics YOLOv5," 2020. [Online]. Available: https://github.com/ultralytics/yolov5.

[16] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8759–8768.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE transactions on pattern analysis and machine intelligence,* vol. 37, no. 9, pp. 1904–1916, 2015.

[18] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[19] Ultralytics, "Ultralytics YOLOv8," 2023. [Online]. Available: https://github.com/ultralytics/ultralytics.

[20] C.-Y. Wang, H.-Y. M. Liao, Y.-H. Wu, P.-Y. Chen, J.-W. Hsieh, and I.-H. Yeh, "CSPNet: A new backbone that can enhance learning capability of CNN," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 390–391.

[21] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6202–6211.

[22] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, "Towards understanding action recognition," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 3192–3199.

[23] O. Köpüklü, X. Wei, and G. Rigoll, "You only watch once: A unified cnn architecture for real-time spatiotemporal action localization," *arXiv preprint arXiv:1911.06644,* 2019.

[24] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems,* vol. 28, 2015.

[25] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.

[26] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, 2009: Ieee, pp. 248–255.

[27] T.-Y. Lin *et al.*, "Microsoft coco: Common objects in context," in *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, 2014: Springer, pp. 740–755.

[28] R. Girshick, I. Radosavovic, G. Gkioxari, P. Dollar, and K. He, "Detectron," 2011. [Online]. Available: https://github.com/facebookresearch/detectron.

[29] "奧奇牌 TW-2700-E7K 測驗考試型雙輪發球機." [Online]. Available: https://www.ttta.com.tw/product.aspx?id=ff917306ee7f463a8d3ad7ab15eac9f2&cid=90b746ba18f64c15a8f21a59cbb7fd14&pcid=.