

Données synthétiques et vision par ordinateur : Vers une reconnaissance robuste des actions en boxe

par

Yosr TAKOUTI

MÉMOIRE PRÉSENTÉ À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
COMME EXIGENCE PARTIELLE À L'OBTENTION DE LA MAÎTRISE EN
GÉNIE DES TECHNOLOGIES DE L'INFORMATION
AVEC MÉMOIRE
M. Sc. A.

MONTRÉAL, LE 12 DÉCEMBRE 2025

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC



Yosr Takouti, 2025



Cette licence Creative Commons signifie qu'il est permis de diffuser, d'imprimer ou de sauvegarder sur un autre support une partie ou la totalité de cette oeuvre à condition de mentionner l'auteur, que ces utilisations soient faites à des fins non commerciales et que le contenu de l'oeuvre n'ait pas été modifié.

PRÉSENTATION DU JURY

CE MÉMOIRE A ÉTÉ ÉVALUÉ

PAR UN JURY COMPOSÉ DE:

M. Sheldon Andrews, directeur de mémoire

Département de génie logiciel et des technologies de l'information à l'École de technologie supérieure

M. David Labbé, codirecteur

Département de génie logiciel et des technologies de l'information à l'École de technologie supérieure

M. Matthew Toews, président du jury

Département de génie des systèmes à l'École de technologie supérieure

M. Christian Desrosiers, membre du jury

Département de génie logiciel et des technologies de l'information à l'École de technologie supérieure

IL A FAIT L'OBJET D'UNE SOUTENANCE DEVANT JURY ET PUBLIC

LE 24 NOVEMBRE 2025

À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

REMERCIEMENTS

Je tiens à exprimer ma profonde gratitude à mes directeurs, **Sheldon Andrews** et **David Labbé**, pour leur encadrement exceptionnel, leurs conseils avisés et leur soutien indéfectible tout au long de ce projet. Leur disponibilité, leur bienveillance et leur capacité à m'accompagner bien au-delà du rôle traditionnel de professeurs ont été pour moi une véritable source d'inspiration. Dans cette aventure vécue loin de ma famille, ils ont représenté un repère essentiel, une présence rassurante et un véritable soutien humain, presque comme une famille de substitution. Je souhaite également adresser mes remerciements les plus sincères au jury, pour le temps qu'il a consacré à l'évaluation de mon travail, pour la qualité de ses commentaires et pour la bienveillance de ses échanges. Ma reconnaissance s'étend aussi à mes collaborateurs du projet de boxe à l'**INS Québec**, en particulier **Thomas Romeas** et **Mathieu Charbonneau**, dont l'expertise, les discussions constructives et l'implication constante ont contribué de manière déterminante à la réussite de ce travail. Leur collaboration a permis d'établir un lien précieux entre la recherche scientifique et la réalité du terrain sportif, enrichissant profondément la portée et la pertinence de ce projet. Je pense également avec reconnaissance à mes collègues et amis de l'équipe de recherche, qui ont su créer un environnement à la fois stimulant, bienveillant et solidaire. Travailler à leurs côtés a été un véritable plaisir : ensemble, nous avons partagé des doutes, des idées, des réussites et de nombreux moments d'encouragement mutuel. À mes amis proches, ici au Canada, en Tunisie et en Europe, merci du fond du cœur. Vos messages, vos appels, vos mots d'encouragement et vos fous rires ont souvent été ce dont j'avais besoin pour tenir bon. Enfin, ma gratitude la plus profonde va à ma famille, pour son amour inconditionnel, son soutien sans faille et sa foi en moi, même à distance. Leur présence dans ma vie est ma plus grande force, et rien de tout cela n'aurait été possible sans eux. Ce parcours n'a pas toujours été facile : il a été fait de défis, de remises en question et de beaucoup d'apprentissage. Mais grâce à toutes ces personnes extraordinaires, j'ai pu surmonter les obstacles, grandir et aller au bout de ce rêve. Aujourd'hui, je ressens une immense fierté et une profonde reconnaissance envers toutes celles et ceux qui ont rendu ce chemin possible.

Données synthétiques et vision par ordinateur : Vers une reconnaissance robuste des actions en boxe

Yosr TAKOUTI

RÉSUMÉ

L'analyse automatisée des performances en boxe soulève des défis scientifiques majeurs en raison du manque de jeux de données annotés de haute qualité, de la complexité des interactions corporelles, de la variabilité des postures et des effets d'occlusion dans les vidéos réelles. La collecte manuelle de telles annotations étant coûteuse, chronophage et sujette à l'erreur humaine, ce mémoire propose une approche alternative fondée sur la génération de données synthétiques. Nous concevons un pipeline complet et modulable permettant de simuler des scènes de boxe réalistes en 3D, en combinant des modèles corporels paramétriques (SMPL), des animations articulaires issues de captures de mouvement et des environnements procéduraux configurés selon plusieurs styles (salle d'entraînement, compétition olympique, Las Vegas, etc.). Ce pipeline intègre également une méthode d'annotation automatique multi-niveaux, enrichie par une annotation manuelle complémentaire et un processus de validation, pour capturer les actions, postures biomécaniques et résultats associés. Grâce à la diversité des caméras virtuelles, des conditions lumineuses et des morphologies corporelles simulées, notre jeu de données synthétique couvre un large éventail de configurations visuelles. Les résultats expérimentaux montrent que les modèles d'apprentissage profond entraînés sur ces données atteignent des performances compétitives en reconnaissance d'action et en estimation de posture, y compris dans des contextes réels. Cette approche ouvre des perspectives innovantes pour le développement de systèmes intelligents appliqués à l'analyse de performance en sports de combat et à la modélisation biomécanique assistée par l'intelligence artificielle.

Mots-clés: Infographie, animation de personnages, ensemble de données synthétiques, analyse de la boxe, classificateur ML

Synthetic data and computer vision : Towards robust boxing action recognition

Yosr TAKOUTI

ABSTRACT

Automated boxing performance analysis presents significant challenges due to the scarcity of high-quality annotated datasets, the complexity of body-to-body interactions, the variability of postures, and frequent occlusions in real-world footage. Manual annotation is time-consuming, error-prone, and difficult to scale, motivating the use of synthetic data generation as an alternative. In this thesis, we present a modular and reproducible pipeline for creating photorealistic 3D boxing scenes, leveraging parametric human models (SMPL), motion capture-driven animations, and procedurally generated environments representing diverse contexts (e.g., training gym, Olympic competition, Las Vegas events). The pipeline incorporates a multi-level automatic annotation system, supported by manual verification, to capture detailed information on actions, biomechanical postures, and outcome categories. A dedicated web-based annotation tool offers synchronized multi-view playback, temporal segmentation, and hierarchical labeling. The rendering engine generates large-scale datasets with rich variations in camera angles, lighting conditions, and body morphologies, enabling robust training samples. Experimental evaluations demonstrate that deep learning models trained on these synthetic data achieve competitive accuracy in action recognition and pose estimation, even when applied to real footage. This work highlights the potential of synthetic datasets in advancing computer vision systems for combat sports performance analysis and AI-assisted biomechanical modeling.

Keywords: Computer graphics, character animation, synthetic dataset, boxing analysis, ML classifier

TABLE DES MATIÈRES

	Page
INTRODUCTION	1
CHAPITRE 1 REVUE DE LITTÉRATURE	5
1.1 Introduction	5
1.2 Concepts fondamentaux et cadre théorique	5
1.2.1 Formats et représentations numériques du mouvement	5
1.2.2 Modèles corporels paramétriques : SMPL et extensions	6
1.2.3 Vision par ordinateur et analyse du mouvement	6
1.2.4 Estimation de pose humaine	7
1.2.5 Reconnaissance d'actions spatio-temporelles	7
1.2.6 Génération et utilisation de données synthétiques	8
1.2.7 Retargeting et adaptation du mouvement	8
1.2.8 Cadre biomécanique et analyse de performance	9
1.3 Estimation de Poses 2D et 3D	9
1.3.1 Définition et enjeux de l'estimation de poses	9
1.3.2 Évolution des approches classiques à l'apprentissage profond	10
1.3.3 Architectures modernes et innovations récentes	16
1.3.4 Gestion des scénarios multi-personnes	17
1.4 Reconnaissance d'actions spatio-temporelles	19
1.4.1 Fondements Conceptuels	19
1.4.2 Évolution des descripteurs manuels aux méthodes profondes	20
1.4.3 Architectures modernes et innovations récentes	21
1.4.4 Gestion des séquences temporelles et modélisation dynamique	23
1.4.5 Modèle de référence : ProtoGCN	24
1.5 Génération des données synthétiques	25
1.5.1 Motivations et Enjeux	26
1.5.2 Modélisation du corps humain : SMPL et ses extensions	26
1.5.3 Systèmes de capture de mouvement et retargeting	28
1.5.4 Randomisation et adaptation de domaine	29
1.5.5 Applications spécifiques à la boxe	30
1.5.5.1 Détection automatique de coups	31
1.5.5.2 Analyse de performance tactique	32
1.6 Synthèse et positionnement de la recherche	33
CHAPITRE 2 MÉTHODOLOGIE ET PIPELINE DE GÉNÉRATION DE FONDÉES SYNTHÉTIQUES	35
2.1 Introduction	35
2.2 Architecture générale du système	35
2.2.1 Conception modulaire et flux de données	36
2.2.2 Gestion des paramètres	37

2.2.2.1	Styles de rings de boxe	37
2.2.2.2	Caméras et éclairages	42
2.2.3	Intégration avec Blender open source	44
2.2.4	Harmonisation des Systèmes de Coordonnées entre Blender et les Données de Capture Réelle	44
2.3	Modélisation et animation des boxeurs virtuels	45
2.3.1	Intégration du modèle SMPL	45
2.3.2	Habillage et apparence des boxeurs virtuels	47
2.3.3	Système de rigging et contrôle d'animation	49
2.3.4	Capture et retargeting de mouvements de boxe avec Rokoko	50
2.3.5	Simulation d'environnements de combat	50
2.3.6	Génération des boxeurs virtuels	52
2.3.7	Pipeline de rendu	53
2.3.8	Stratégies d'optimisation computationnelle	54
2.4	Génération des annotations de vidéos	55
2.4.1	Notre outil d'annotation	55
2.4.2	Annotation des vidéos de boxe	59
2.5	Analyse comparative des jeux de données	61
CHAPITRE 3 MODÈLES ET ARCHITECTURES IMPLEMENTÉS		63
3.1	Introduction	63
3.2	Pipeline Général d'Estimation de Pose et de Classification des Actions	63
3.3	Détection et Estimation de la Pose des boxeurs	65
3.3.1	Architecture ViTPose et Adaptations Spécifiques	65
3.3.2	Méthode d'Appariement 3D/2D	66
3.3.3	Gestion des Occlusions et Interactions	66
3.3.4	Détection des personnes dans le jeu de données réel	67
3.3.5	Contrainte spatiale : détection de la zone du ring	68
3.3.6	Association et suivi des boxeurs	70
3.3.7	Sorties générées	71
3.4	Classification des Actions	71
CHAPITRE 4 RÉSULTATS ET ANALYSE DE PERFORMANCE		73
4.1	Introduction	73
4.2	Prétraitement des données pour le modèle de base	73
4.3	Optimisation des hyperparamètres	75
4.3.1	Recherche par grille.	75
4.3.2	Optimization via l'intégration d'Optuna	76
4.3.3	Résultats d'entraînement et de l'optimisation sur le jeu de données de boxe	77
4.4	Performances et Analyse Comparative	80
4.4.1	Entraînement sur le jeu de données réel de boxe	83
4.4.2	Entraînement sur le jeu de données réel et synthétique total	87
4.4.2.1	Stratification des données pour l'entraînement	87

	4.4.2.2	Optimisation des hyperparamètres et entraînement	88
	4.4.2.3	Analyse des résultats	89
4.5		Analyse des Résultats	90
	4.5.1	Analyse Comparative avec l'État de l'Art	90
	4.5.2	Performances et Analyse Comparative	92
	4.5.3	Comparaison des performances par classe : Réel vs Synthétique vs Réel + Synthétique	94
	4.5.4	Évaluation de l'approche multi-classes à fenêtre temporelle fixe	95
	4.5.5	Discussion	98
	4.5.6	Effet des Données Synthétiques sur les Performances	100
CONCLUSION ET RECOMMANDATIONS			103
ANNEXE I	GLOSSAIRE		105
ANNEXE II	ACTIONS DE BOXE EN DESCRIPTION		111
ANNEXE III	EXEMPLES D'ACTIONS DE BOXE EN IMAGES		113
ANNEXE IV	LES PIPELINES DE LA CRÉATION DE JEU DE DONNÉES SYNTHÉTIQUE		119
ANNEXE V	EXEMPLES DE CLASSIFICATION DES ACTIONS DE BOXE EN IMAGES		121
BIBLIOGRAPHIE			127

LISTE DES TABLEAUX

	Page
Tableau 4.1	Comparaison des précisions pour différentes actions de boxe entre le modèle de Seminario et notre modèle 81
Tableau 4.2	Comparaison des rappels pour différentes actions de boxe entre le modèle de Seminario et notre modèle 81
Tableau 4.3	Comparaison des scores F1 pour différentes actions de boxe entre le modèle de Seminario et notre modèle 82
Tableau 4.4	Comparaison des scores par classe : précision, rappel et F1 de notre modèle, ainsi que le F1-score moyen du modèle de Seminario 82
Tableau 4.5	Résultats de classification sur le jeu de données réel de boxe 84
Tableau 4.6	Résultats de classification sur le jeu de données synthétique correspondant (mêmes vidéos reproduites) 85
Tableau 4.7	Comparaison des performances de ST-GCN et ProtoGCN sur plusieurs jeux de données de reconnaissance d'actions basés sur le squelette, incluant notre jeu de données de boxe 91
Tableau 4.8	Comparaison des performances par classe selon le type de jeu de données : réel, synthétique et combiné (réel + synthétique) 94

LISTE DES FIGURES

	Page
Figure 1.1	Exemple d'image issue de la base LSP illustrant les points-clés corporels 11
Figure 1.2	Cartes thermiques 12
Figure 1.3	Tests de MPII 13
Figure 1.4	Méthodologie MixSTE 15
Figure 1.5	Résultats de MixSTE 15
Figure 1.6	Suivi de la Pose 17
Figure 1.7	OpenPose 18
Figure 1.8	SIP vs STIP 20
Figure 1.9	I3D 22
Figure 1.10	AP par catégorie sur AVA 23
Figure 1.11	ST-GCN 24
Figure 1.12	Architecture globale de ProtoGCN : Un réseau de reconstruction de prototypes est proposé pour transformer la représentation de la topologie du graphe X en une combinaison Z de prototypes apprenables au niveau des articulations fines 25
Figure 1.13	Génération d'un jeu de données synthétique à partir de modèles SMPL et de séquences de motion capture 27
Figure 1.14	SMPL 28
Figure 1.15	Exemple de randomisation structurée par domaine (SDR) 29
Figure 1.16	Boxing Punch Detection with Single Static Camera 31
Figure 1.17	Précision de la classification des coups de poing 32
Figure 2.1	Pipeline de génération des données synthétiques 36
Figure 2.2	Aperçu général de rings de boxe 38

Figure 2.3	Aperçu de ring de boxe dans le style Las Vegas	39
Figure 2.4	Aperçu de ring de boxe dans les Jeux Olympiques	40
Figure 2.5	Aperçu de ring de boxe dans une salle d'entraînement	41
Figure 2.6	Vue d'ensemble du jeu de données synthétiques	42
Figure 2.7	échantillon des textures d'avatars boxeurs	49
Figure 2.8	Pipeline de la création du boxeur virtuel	53
Figure 2.9	Interface de l'outil d'annotation avec vue multi-caméras et timeline synchronisée	60
Figure 2.10	Étiquetage hiérarchique des actions techniques et des segments temporels annotés	61
Figure 3.1	Pipeline général d'estimation de pose et de reconnaissance d'action	65
Figure 3.2	L'effet des paramètres du pipeline de détection dans les vidéos réelles ...	69
Figure 3.3	Analyse du modèle ProtoGCN	72
Figure 4.1	Le mappage des 20 actions en 8 classes	75
Figure 4.2	Les résultats d'entraînement pour la classification sur le jeu de données synthétique de base	78
Figure 4.3	Performance du modèle par classe pour la classification avec ProtoGCN avec les jeux de données synthétiques	79
Figure 4.4	Exemple de résultat de classification des actions de boxe : cas de garde neutre vs directe	83
Figure 4.5	Résultat de classification sur le jeu de données réel	86
Figure 4.6	Résultat de classification sur le jeu de données synthétique similaire à la réelle	86
Figure 4.7	Modèle de mouvement temporel des actions de boxe	87
Figure 4.8	Comparaison entre ProtoGCN et ST-GCN	90
Figure 4.9	Résultat d'entraînement final pour la classification avec ProtoGCN avec les jeux de données réels et synthétiques	93

Figure 4.10	Les résultats des métriques par classe pour la classification avec ProtoGCN avec les jeux de données réels et synthétiques	93
Figure 4.11	Les métriques par classe	96
Figure 4.12	Matrice de confusion normalisée pour les classes	97
Figure 4.13	Exemple de détection d'actions dans une vidéo synthétique	98

INTRODUCTION

Dans le contexte du sport de haut niveau, l'analyse quantitative des performances constitue un enjeu majeur pour l'optimisation de l'entraînement et l'amélioration des résultats compétitifs. En boxe, cette analyse vise à identifier et quantifier les techniques offensives et défensives déployées par les athlètes, permettant une évaluation objective de leurs capacités tactiques et techniques. Traditionnellement, ce processus repose sur l'observation manuelle par des experts, l'annotation chronophage de séquences vidéo, et l'interprétation subjective de performances, ce qui limite considérablement son accessibilité et sa reproductibilité à grande échelle. L'avènement de la vision par ordinateur et des techniques d'apprentissage profond a ouvert de nouvelles perspectives pour l'automatisation de ces analyses. Cependant, l'entraînement de modèles performants demeure confronté à un défi fondamental : la rareté et la complexité de l'acquisition de données annotées de qualité. L'objectif ultime consiste à fournir aux entraîneurs et aux athlètes des outils objectifs d'évaluation permettant d'optimiser les stratégies d'entraînement et d'améliorer les performances compétitives. En contexte réel, la collecte de vidéos exploitables de combats de boxe présente plusieurs contraintes majeures. Premièrement, l'accès aux athlètes de haut niveau est limité en raison de leurs emplois du temps chargés et des restrictions liées à la confidentialité des entraînements. Deuxièmement, les conditions de capture sont souvent sous-optimales : éclairage variable, angles de caméra restreints, occlusions fréquentes dues aux mouvements rapides et aux interactions entre boxeurs. Troisièmement, les coûts élevés des équipements de suivi professionnel et la complexité de leur mise en œuvre constituent des barrières significatives. Enfin, la distribution déséquilibrée des classes d'actions pose un problème particulier, notamment pour les techniques rares mais tactiquement décisives. Face à ces limitations, la génération de données synthétiques s'impose comme une solution innovante et prometteuse. Cette approche permet de créer des jeux de données massifs, diversifiés et parfaitement équilibrés, tout en contournant les contraintes pratiques et éthiques liées à la collecte de données réelles. En simulant des environnements virtuels, des mouvements corporels

précis et des situations de combat variées, il devient possible de constituer des jeux de données annotés, réduisant considérablement les coûts et le temps associés à l'acquisition de données tout en améliorant la reproductibilité et la rigueur des analyses (Sengupta, Budvytis & Cipolla, 2020). Cette thèse s'inscrit dans cette dynamique et propose un cadre méthodologique original pour l'analyse de performance en boxe à partir de données synthétiques. Notre approche combine la capture de mouvement, la modélisation 3D avancée et les réseaux de neurones profonds pour permettre l'estimation de poses 2D et la reconnaissance d'actions spatio-temporelles.

Dans ce cadre, cette thèse propose une approche innovante, dont les contributions principales sont :

1. un **nouveau pipeline méthodologique original** pour la génération de données synthétiques de boxe visant la diversité biomécanique
2. une **base de données de boxe de haute qualité** qui comprend nombreuses vidéos synthétiques diversifiées
3. un **système d'analyse et de validation** reposant sur des derniers état de l'art architectures d'apprentissage profond capable de détecter et classer les actions de boxe

Ce travail est réalisé en collaboration étroite avec l'Institut national du sport du Québec (INS Québec), un centre d'entraînement de référence pour les boxeurs d'élite, qui combine des activités de recherche, d'innovation et de développement technologique en vue d'optimiser la performance sportive. Ce partenariat avec les experts et athlètes de l'INS Québec permet d'accéder à une connaissance approfondie des besoins technologiques propres aux données d'entraînement. Ces données sont essentielles pour développer, calibrer et évaluer notre méthode de création des données synthétiques et de réaliser la classification des actions de boxe.

Cette thèse s’articule en quatre chapitres complémentaires. Le **chapitre 1** présente un état de l’art approfondi portant sur l’estimation de la pose 2D, la reconnaissance d’actions et la génération de données synthétiques, en mettant l’accent sur les approches récentes. Le **chapitre 2** détaille notre pipeline de génération de scènes de boxe synthétiques, depuis la capture de mouvement jusqu’au rendu final. Le **chapitre 3** expose les modèles entraînés, les métriques adoptées ainsi que les résultats obtenus en classification et en analyse de performance. Le **chapitre 4** propose une analyse quantitative et qualitative des résultats en expliquant la démarche de travail avec une discussion sur la classification et la prédiction des actions de boxe. Enfin, la conclusion synthétise les contributions, en discutant des limites rencontrées et en ouvrant sur des perspectives d’extension de cette approche à d’autres sports de combat ou domaines d’analyse du mouvement humain.

CHAPITRE 1

REVUE DE LITTÉRATURE

1.1 Introduction

L'analyse automatique des performances sportives représente un domaine de recherche en pleine expansion, situé à la convergence de la vision par ordinateur, de l'apprentissage automatique et des sciences du sport. Ce premier chapitre propose un état de l'art approfondi des domaines scientifiques et techniques qui sous-tendent notre approche. Nous examinerons d'abord les fondements de l'estimation de poses 2D, une technique centrale pour la compréhension des mouvements humains. Nous explorerons ensuite les méthodes de reconnaissance d'actions spatio-temporelles, essentielles pour l'identification automatique des techniques de boxe. Enfin, nous analyserons les approches de génération de données synthétiques, en mettant l'accent sur leurs applications dans le domaine du sport et de l'analyse du mouvement humain.

1.2 Concepts fondamentaux et cadre théorique

Cette section présente les notions essentielles qui forment la base de notre recherche. Elle clarifie les termes techniques et les concepts mobilisés dans ce mémoire afin de fournir un cadre théorique cohérent à l'analyse des performances en boxe par vision par ordinateur.

1.2.1 Formats et représentations numériques du mouvement

L'un des premiers éléments à considérer est la représentation numérique des mouvements humains. Deux formats standards sont largement utilisés dans l'industrie et la recherche. Le format FBX (FilmBoX), développé par Autodesk, permet l'échange de modèles 3D, d'animations et de métadonnées entre divers logiciels de modélisation et de rendu. Il est devenu un standard dans l'animation et la capture de mouvement, grâce à sa capacité à stocker simultanément géométrie, squelette, matériaux et animation. Le format BVH (BioVision Hierarchy), plus ancien, organise les articulations humaines sous forme hiérarchique et encode directement les rotations

de chaque articulation au fil du temps. Ces formats constituent la base des pipelines modernes de motion capture et servent de point d'entrée pour l'intégration avec des modèles paramétriques comme SMPL.

1.2.2 Modèles corporels paramétriques : SMPL et extensions

Le modèle SMPL (Skinned Multi-Person Linear Model) (Loper, Mahmood, Romero, Pons-Moll & Black, 2015) est un modèle statistique du corps humain qui décrit la surface corporelle par un maillage triangulaire déformable. Il se distingue par sa capacité à générer des avatars réalistes contrôlés par deux ensembles de paramètres : les paramètres de forme, qui capturent la variabilité morphologique des individus, et les paramètres de pose, qui définissent la configuration articulaire. Ce modèle a largement contribué à l'avancée des recherches en vision par ordinateur et en infographie car il permet de produire à la fois des représentations 3D réalistes et des annotations de pose 2D/3D cohérentes. Ses extensions, comme SMPL-X (Pavlakos *et al.*, 2019) ou STAR (Osman, Bolkart & Black, 2020), ajoutent la prise en compte des mains, du visage et des déformations musculaires pour un rendu encore plus complet.

1.2.3 Vision par ordinateur et analyse du mouvement

La vision par ordinateur est la discipline qui cherche à automatiser l'interprétation des images et vidéos. Dans le cadre de l'analyse du mouvement, son objectif est de détecter, suivre et comprendre la dynamique corporelle humaine. L'estimation de pose, en 2D ou en 3D, est au cœur de cette problématique : elle consiste à localiser les articulations clés du squelette humain à partir de données visuelles (Cao, Hidalgo, Wei & Sheikh, 2019). Cette information est ensuite utilisée pour analyser la cinématique des mouvements et inférer des actions. En sport, elle permet de transformer une séquence vidéo brute en représentation biomécanique objective.

1.2.4 Estimation de pose humaine

L'estimation de pose 2D a connu un essor avec des architectures comme OpenPose (Cao *et al.*, 2019), HRNet (Sun, Xiao, Liu & Wang, 2019) et plus récemment ViTPose (Xu *et al.*, 2022), qui exploitent les Transformers pour capturer des dépendances spatiales et temporelles complexes. L'estimation de pose 3D repose quant à elle sur des approches qui reconstruisent les coordonnées spatiales des articulations à partir de séquences multi-vues ou monoculaires (Martinez, Hossain, Romero & Little, 2017; Pavllo, Feichtenhofer, Grangier & Auli, 2019). Ces techniques constituent un socle essentiel pour relier les observations visuelles au cadre biomécanique du sport.

1.2.5 Reconnaissance d'actions spatio-temporelles

La reconnaissance d'actions vise à identifier automatiquement les gestes exécutés par un athlète à partir d'une séquence vidéo, tels qu'un jab, un crochet ou un blocage. Les approches classiques reposaient sur des descripteurs manuels comme les STIP (Laptev, 2005) ou les HOG/HOF (Dalal, Triggs & Schmid, 2006), mais ces méthodes demeuraient limitées face aux variations d'apparence, d'éclairage et de perspective. L'émergence des réseaux profonds tridimensionnels, tels que I3D (Carreira & Zisserman, 2017) ou SlowFast (Feichtenhofer, Fan, Malik & He, 2019), a permis d'extraire conjointement des caractéristiques spatiales et temporelles. Cependant, ces modèles volumétriques, bien qu'efficaces pour la classification d'actions génériques, restent coûteux en ressources et peu adaptés à la modélisation explicite des articulations humaines. Pour surmonter ces limites, les approches squelettiques comme le *Spatial-Temporal Graph Convolutional Network* (ST-GCN) (Yan, Xiong & Lin, 2018) ont introduit une représentation en graphe du corps humain, où chaque articulation constitue un nœud et les liens anatomiques forment les arêtes. Ce paradigme permet de modéliser les dépendances spatio-temporelles entre articulations, offrant une représentation plus robuste et biomécaniquement cohérente. Dans cette continuité, ProtoGCN (Liu *et al.*, 2025) propose une modélisation par prototypes articulaires, capable de distinguer des gestes visuellement proches et d'améliorer la généralisation sur des mouvements rapides ou partiellement occultés. Cette approche s'avère particulièrement adaptée

à la boxe, où la précision et la dynamique des postures jouent un rôle central, et elle justifie le choix méthodologique adopté dans ce travail.

1.2.6 Génération et utilisation de données synthétiques

Un défi central de l'analyse de la boxe réside dans le manque de données annotées de haute qualité. La génération de données synthétiques est apparue comme une solution innovante pour pallier cette rareté (Varol *et al.*, 2017; Sengupta *et al.*, 2020). Elle repose sur l'utilisation de modèles paramétriques tels que SMPL, de systèmes de capture de mouvement et de moteurs de rendu comme Blender ou Unreal Engine pour simuler des scènes de combat. Plus récemment, le jeu de données BEDLAM (Black, Patel, Tesch & Yang, 2023) a marqué une avancée majeure en combinant capture de mouvement, rendu photoréaliste et diversité morphologique à grande échelle, offrant un cadre de référence pour l'apprentissage sur corps humains virtuels. Ces données synthétiques présentent plusieurs avantages : contrôle total sur la diversité des morphologies, des environnements et des conditions visuelles ; génération automatique d'annotations et équilibrage des classes d'actions rares. Elles permettent ainsi de renforcer l'entraînement des modèles d'apprentissage profond et d'améliorer leur robustesse face aux variations réelles observées en compétition.

1.2.7 Retargeting et adaptation du mouvement

Le motion retargeting consiste à transférer un mouvement capturé d'un individu vers un modèle virtuel de morphologie différente (Gleicher, 1998; Villegas *et al.*, 2018). Cette technique est indispensable pour adapter les séquences de motion capture aux modèles paramétriques tels que SMPL, tout en préservant la cohérence biomécanique et la fluidité du geste. Les approches récentes combinent optimisation sous contraintes et apprentissage profond afin d'assurer un transfert réaliste comme les réseaux neuronaux à structure squelettique (Aberman *et al.*, 2020) ou les modèles guidés par la pose améliorent la fidélité et la généralisabilité du mouvement. Dans ce travail, le retargeting constitue une étape essentielle du pipeline, permettant de transférer les mouvements réels des boxeurs vers des avatars SMPL standardisés. Cette normalisation

garantit la cohérence entre les données réelles et synthétiques pour l'apprentissage profond et l'analyse de performance.

1.2.8 Cadre biomécanique et analyse de performance

L'analyse de performance sportive repose sur la traduction des données visuelles en indicateurs pertinents pour l'entraînement. Les variables biomécaniques extraites (vitesse, amplitude, équilibre, posture) sont interprétées à la lumière des sciences du sport (Bartlett, 2014). En boxe, elles permettent de mesurer l'efficacité des actions, la réactivité défensive, la qualité technique et la stratégie globale de l'athlète. L'articulation entre vision par ordinateur, modélisation biomécanique et sciences du sport constitue ainsi le cadre théorique global de ce mémoire.

1.3 Estimation de Poses 2D et 3D

Cette section présente les principes essentiels depuis les approches géométriques classiques jusqu'aux architectures modernes à base de réseaux de neurones profonds et de transformeurs. Elle met en évidence les enjeux spécifiques liés aux scénarios multi-personnes, tels que les combats de boxe, où la cohérence temporelle et la précision des articulations jouent un rôle déterminant pour une interprétation fiable des actions.

1.3.1 Définition et enjeux de l'estimation de poses

L'estimation de poses 2D consiste à localiser et à identifier les articulations clés du corps humain dans des images ou des séquences vidéo. Cette tâche fondamentale de la vision par ordinateur vise à extraire une représentation structurée du squelette humain sous la forme d'un ensemble de points d'intérêt connectés par des liens anatomiques. Dans le contexte de l'analyse sportive, et particulièrement en boxe, cette information squelettique constitue la base pour comprendre et analyser les mouvements des athlètes. La représentation classique d'une pose humaine s'appuie sur un modèle de squelette composé de N articulations, typiquement entre 14 et 25 points selon le niveau de détail souhaité. Chaque articulation est caractérisée par ses coordonnées

bidimensionnelles (x, y) dans l'espace image, accompagnées d'un score de confiance indiquant la fiabilité de la détection. Les articulations sont généralement organisées selon une hiérarchie anatomique, reliant par exemple l'épaule au coude, puis le coude au poignet, formant ainsi une chaîne cinématique cohérente. Dans un cadre tridimensionnel, l'estimation de poses 3D vise à reconstruire les coordonnées spatiales complètes (x, y, z) des articulations à partir d'images 2D. Cette tâche nécessite d'inférer la profondeur à partir d'indices visuels et temporels. Elle est cruciale dans les applications sportives car elle permet de déduire la cinématique corporelle, la posture réelle du corps et les amplitudes des gestes. Toutefois, pour notre pipeline, l'estimation 2D reste privilégiée car elle offre une robustesse supérieure dans les scènes non calibrées et alimente efficacement les modules de reconnaissance d'actions.

1.3.2 Évolution des approches classiques à l'apprentissage profond

Les premières approches d'estimation de poses s'appuyaient sur des méthodes de traitement d'image classiques combinées à des modèles géométriques explicites. Ces techniques, développées dans les années 1990 et 2000, utilisaient des détecteurs de contours, des filtres de Gabor et des modèles de formes déformables pour localiser les parties du corps. L'algorithme des Pictorial Structures, proposé par Fischler et Elschlager, représente un repère important de cette époque. Cette méthode modélise le corps humain comme un ensemble de parties rigides connectées par des joints flexibles, permettant de formuler l'estimation de poses comme un problème d'optimisation graphique. Cependant, ces approches classiques souffraient de limitations importantes. Leur dépendance à des caractéristiques visuelles peu robustes les rendait sensibles aux variations d'apparence et aux conditions d'éclairage. De plus, leur formulation explicite des contraintes anatomiques s'avérait insuffisamment flexible pour capturer la diversité des poses humaines possibles. L'émergence de l'apprentissage profond a révolutionné le domaine de l'estimation de poses à partir de 2014. Les réseaux de neurones convolutionnels (CNN) ont démontré leur capacité à apprendre automatiquement des représentations visuelles robustes directement à partir des données d'entraînement. (Toshev & Szegedy, 2014) constitue l'une des premières applications réussies des CNN à l'estimation de poses. Cette approche formule

le problème comme une tâche de régression directe, où le réseau prédit les coordonnées des articulations à partir de l'image d'entrée. La Figure 1.1 illustre des exemples issus de la base LSP (Leeds Sports Pose) présentant des résultats visuels de l'estimation de poses obtenus à partir de cette approche. On y observe la capacité du réseau à localiser les points clés corporels sur divers types de postures et d'activités sportives mettant en évidence la généralisation du modèle aux variations d'échelle de vêtements et de conditions d'arrière-plan. Cette illustration montre ainsi l'un des premiers succès du deep learning dans la modélisation du corps humain en 2D. La Figure 1.2 illustre le principe de cette approche : chaque carte thermique encode



Figure 1.1 Exemple d'image issue de la base LSP illustrant les points-clés corporels (Toshev & Szegedy, 2014)

la localisation probable d'un point-clé spécifique, et l'estimation finale de la pose est obtenue en sélectionnant le maximum d'activation sur chaque carte. Cette représentation probabiliste offre une meilleure robustesse aux variations d'apparence, aux occultations partielles et aux conditions d'éclairage, marquant une avancée majeure dans l'évolution des architectures CNN pour l'estimation de poses humaines. Rapidement, des architectures plus sophistiquées ont

émergé, exploitant les cartes de chaleur (heatmaps) plutôt que la régression directe. Cette représentation intermédiaire, popularisée par (Newell, Yang & Deng, 2016), offre plusieurs avantages. D’une part, elle transforme le problème de régression en un problème de classification dense, généralement plus stable à optimiser. D’autre part, elle fournit une information spatiale riche qui facilite la localisation précise des articulations même en présence d’ambiguïtés locales. La Figure 1.3 illustre plusieurs exemples de sorties obtenues sur l’ensemble de tests du jeu de données MPII l’un des ensembles de référence les plus utilisés pour l’évaluation de l’estimation de poses 2D. On y observe la capacité du modèle à généraliser à une grande variété de scènes, d’activités et d’angles de vue tout en maintenant une cohérence anatomique entre les points clés détectés. Ces résultats témoignent de la robustesse du réseau face aux variations d’apparence aux interactions entre individus et aux occlusions partielles, confirmant la pertinence des architectures convolutionnelles empilées pour l’estimation de pose humaine en conditions réelles. Un travail marquant dans ce contexte est proposé par Zhang et al. avec *MixSTE* (Seq2seq Mixed

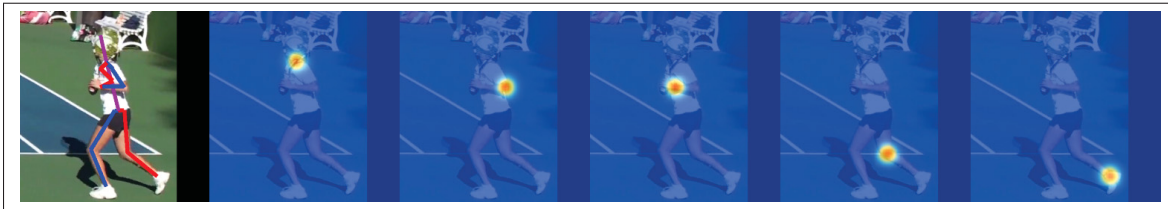


Figure 1.2 À gauche, l’estimation finale de la pose fournie par les activations maximales sur chaque carte thermique (Newell *et al.*, 2016)

Spatio-Temporal Encoder) (Zhang, Tu, Yang, Chen & Yuan, 2022). Contrairement aux approches précédentes, qui considéraient les articulations dans toutes les frames de manière globale, MixSTE introduit une modélisation hybride de la corrélation spatio-temporelle. L’architecture alterne deux blocs complémentaires : un *temporal transformer block*, qui capture le mouvement de chaque articulation indépendamment dans le temps, et un *spatial transformer block*, qui apprend les relations spatiales entre articulations. Cette conception permet de mieux représenter les différences de dynamique entre articulations (par exemple, les mouvements rapides des bras vs. la stabilité relative du torse), ce qui améliore la précision globale de l’estimation. De plus, le

modélisation spatio-temporelle séquentielle d'articulations pour l'estimation 3D à partir de séquences monoculaires, sans intégrer de contraintes physiques ou de dynamique multi-personne. Notre pipeline, bien que conceptuellement inspiré de la représentation séquentielle de MixSTE, s'aligne davantage sur l'approche de Feiz et al., en privilégiant une estimation 2D top-down (ViTPose) et une reconstruction 3D implicite fondée sur des contraintes cinématiques. Ce choix offre une meilleure compatibilité avec nos vidéos réelles et synthétiques de boxe, où les caméras ne sont pas calibrées et les interactions sont rapides et partiellement occultées. Ainsi, MixSTE est discuté ici pour son apport conceptuel dans la modélisation temporelle, tandis que l'approche de Feiz et al. représente l'état de l'art le plus pertinent et directement relié à notre pipeline de recherche. La Figure 1.4 illustre le schéma de cette architecture, où les blocs spatiaux et temporels sont intercalés pour assurer un encodage optimal des corrélations spatio-temporelles. Les coordonnées 2D des points clés sont d'abord enrichies par un encodage de position, puis traitées par une série de modules Transformer capables d'apprendre des dépendances longues, tant au sein d'une trame que sur l'ensemble de la séquence vidéo. Cette conception favorise une compréhension plus fine de la structure du mouvement, tout en réduisant la redondance entre représentations spatiales et temporelles. La Figure 1.5 présente plusieurs exemples de résultats obtenus par MixSTE sur différentes séquences vidéo. On y observe la correspondance entre les images d'entrée et les reconstructions 3D générées, démontrant la capacité du modèle à capturer fidèlement la structure corporelle et la trajectoire des articulations dans l'espace tridimensionnel. Les exemples incluent des actions variées telles que la danse ou le saut, illustrant la robustesse du modèle face à la diversité des postures, des vitesses de mouvement et des angles de vue.

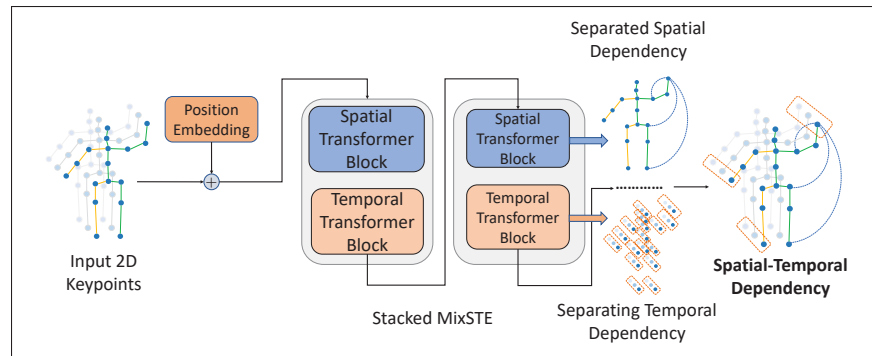


Figure 1.4 L'architecture MixSTE des corrélations spatio-temporelles (Zhang *et al.*, 2022)

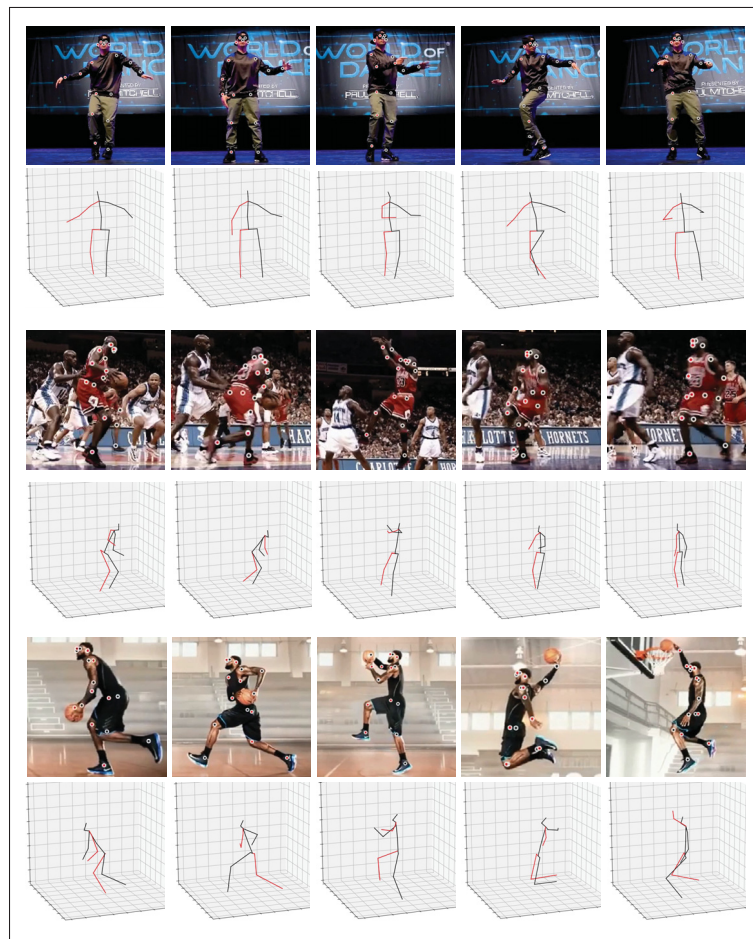


Figure 1.5 Exemple de résultats obtenus par MixSTE sur des séquences vidéo, montrant la précision accrue de l'estimation 3D (Zhang *et al.*, 2022)

1.3.3 Architectures modernes et innovations récentes

Les architectures contemporaines d'estimation de poses se caractérisent par leur complication croissante et leur spécialisation pour différents contextes d'application. L'architecture Stacked Hourglass utilise une succession de modules en forme de sablier qui combinent des opérations de sous-échantillonnage et de sur-échantillonnage. Cette structure permet de capturer des informations à différentes échelles spatiales, essentielle pour localiser précisément des articulations de tailles variables. (Xiao, Wu & Wei, 2018), simplifie cette approche en utilisant une architecture plus directe basée sur ResNet. Malgré sa simplicité apparente, cette méthode atteint des performances remarquables en s'appuyant sur des techniques d'augmentation de données sophistiquées et un entraînement soigneusement optimisé. Cette approche démontre qu'une architecture bien conçue peut rivaliser avec des modèles plus complexes. HRNet introduit une innovation architecturale majeure en maintenant des représentations haute résolution tout au long du processus de traitement. Contrairement aux approches classiques qui réduisent progressivement la résolution spatiale avant de la restaurer, HRNet conserve plusieurs branches parallèles à différentes résolutions, permettant un échange d'information multi-échelle continu. Cette approche s'avère particulièrement efficace pour la localisation précise d'articulations fines. L'émergence des Transformers dans le domaine de la vision par ordinateur a également influencé l'estimation de poses. ViTPose adapte l'architecture Vision Transformer au problème de l'estimation de poses, exploitant les mécanismes d'attention pour capturer les dépendances à long terme entre les différentes parties du corps. Cette approche montre des résultats prometteurs, particulièrement pour gérer les occlusions et les poses complexes. Elle constitue la base de notre pipeline pour l'estimation 2D des boxeurs, avant la phase de reconnaissance d'actions. La Figure 1.6 présente le suivi de la pose basé sur le flux optique qui illustre le processus de propagation et d'association des poses à travers les images successives d'une séquence vidéo. À partir des résultats des trames précédentes, le modèle génère des boîtes de flux et calcule la similarité des poses en exploitant les déformations locales induites par le mouvement. Les poses sont ensuite ajustées dynamiquement par décalage spatial avant d'être appariées de manière gloutonne afin de maintenir la cohérence temporelle des identités suivies.

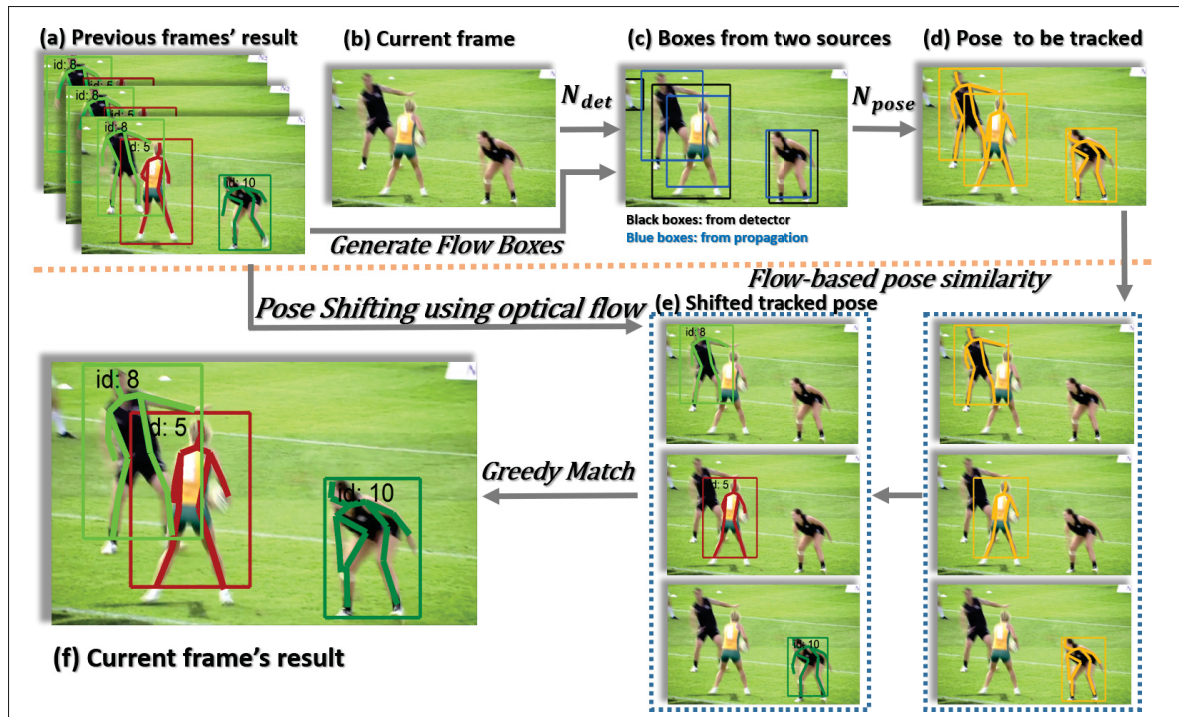


Figure 1.6 Le cadre proposé pour le suivi de la pose basé sur le flux
(Xiao *et al.*, 2018)

1.3.4 Gestion des scénarios multi-personnes

L'estimation de la pose en présence de plusieurs personnes (p. ex., deux boxeurs en interaction) se traite classiquement selon deux familles : top-down et bottom-up.

Top-down : Une détection de personnes est d'abord appliquée pour obtenir un cadre englobant par individu ; un estimateur mono-personne est ensuite exécuté sur chaque recadrage. Cette stratégie, utilisée par les modèles récents à base de transformeurs comme ViTPose, permet d'optimiser la précision par personne et domine aujourd'hui les tableaux COCO en termes d'AP (au prix d'un coût qui croît avec le nombre de personnes). Dans notre cas d'usage (deux boxeurs), ce compromis est favorable : la latence reste maîtrisée tandis que la précision des articulations critiques (mains, coudes, tête) est supérieure.

Bottom-up : À l'inverse, les méthodes bottom-up détectent d'abord l'ensemble des articulations visibles dans l'image, puis les regroupent pour reconstruire chaque squelette. OpenPose en est l'exemple fondateur via les Part Affinity Fields (PAFs) qui encodent la connectivité entre

segments. Leur avantage clef est une complexité qui dépend peu du nombre de personnes et une robustesse appréciable dans les scènes très denses ; en contrepartie, la précision par individu est généralement inférieure aux meilleurs modèles top-down récents. La Figure 1.7 présente les résultats produits par OpenPose tels que proposés par Cao et al. (2019), illustrant la détection multi-personnes et la génération des cartes thermiques associées aux articulations. Le modèle repose sur un mécanisme de Part Affinity Fields (PAF), qui encode les relations directionnelles entre paires de points-clés pour relier correctement les articulations appartenant à un même individu. Cette approche permet d'identifier simultanément plusieurs personnes dans une même scène tout en maintenant la cohérence anatomique de chaque squelette.



Figure 1.7 Résultats heatmaps avec OpenPose
(Cao *et al.*, 2019)

1.4 Reconnaissance d'actions spatio-temporelles

Dans le cadre de l'analyse de la boxe, cette tâche permet de distinguer les techniques offensives et défensives, d'évaluer la qualité d'exécution des mouvements et de quantifier les performances athlétiques. Cette section présente les fondements conceptuels, l'évolution des méthodes de reconnaissance d'actions, les architectures modernes et les stratégies de modélisation temporelle adaptées aux séquences complexes.

1.4.1 Fondements Conceptuels

La reconnaissance d'actions humaines constitue un domaine fondamental de la vision par ordinateur qui vise à identifier et classer automatiquement les activités réalisées par des individus dans des séquences vidéo. Dans le contexte sportif, et particulièrement en boxe, cette tâche consiste à distinguer et catégoriser les différentes techniques offensives et défensives telles que les jabs, crochets, uppercuts, et blocages. Cette classification automatique représente un enjeu majeur pour l'analyse objective des performances et le développement d'outils d'aide à l'entraînement. La complexité de la reconnaissance d'actions réside dans la nature intrinsèquement spatio-temporelle des mouvements humains. Contrairement à la classification d'images statiques, la compréhension d'une action nécessite l'analyse de séquences temporelles, où l'information pertinente se distribue à travers multiple frames. Cette dimension temporelle introduit des défis supplémentaires : variabilité dans la durée d'exécution des actions, vitesses d'exécution différentes selon les individus, et transitions graduelles entre différentes techniques. La représentation des actions soulève également des questions fondamentales. Une action peut être caractérisée à différents niveaux d'abstraction : des mouvements élémentaires des articulations aux patterns complexes impliquant l'ensemble du corps. En boxe, un crochet du gauche implique une coordination précise entre la rotation du torse, l'extension du bras, et le positionnement des jambes. Cette hiérarchie de mouvements nécessite des approches capables de capturer à la fois les détails fins et les patterns globaux.

1.4.2 Évolution des descripteurs manuels aux méthodes profondes

Les premières approches de reconnaissance d'actions s'appuyaient sur des descripteurs visuels conçus manuellement pour capturer les caractéristiques spatio-temporelles pertinentes. Les Space-Time Interest Points (STIP), (Das Dawn & Shaikh, 2016), étendent le concept de points d'intérêt spatiaux aux volumes spatio-temporels. Ces détecteurs identifient des régions présentant des variations significatives à la fois dans l'espace et le temps, théoriquement associées aux parties les plus informatives des actions. La Figure 1.8 illustre une comparaison visuelle entre la détection de points d'intérêt spatiaux (SIP) et les points d'intérêt spatio-temporels (STIP) comme la première rangée montre les résultats de la détection SIP qui identifie des régions statiques pertinentes dans chaque trame en se basant uniquement sur les caractéristiques spatiales de l'image et la seconde rangée présente les résultats obtenus avec la détection STIP, où les points d'intérêt sont extraits en intégrant également la dimension temporelle c'est-à-dire les variations locales de mouvement entre trames successives.

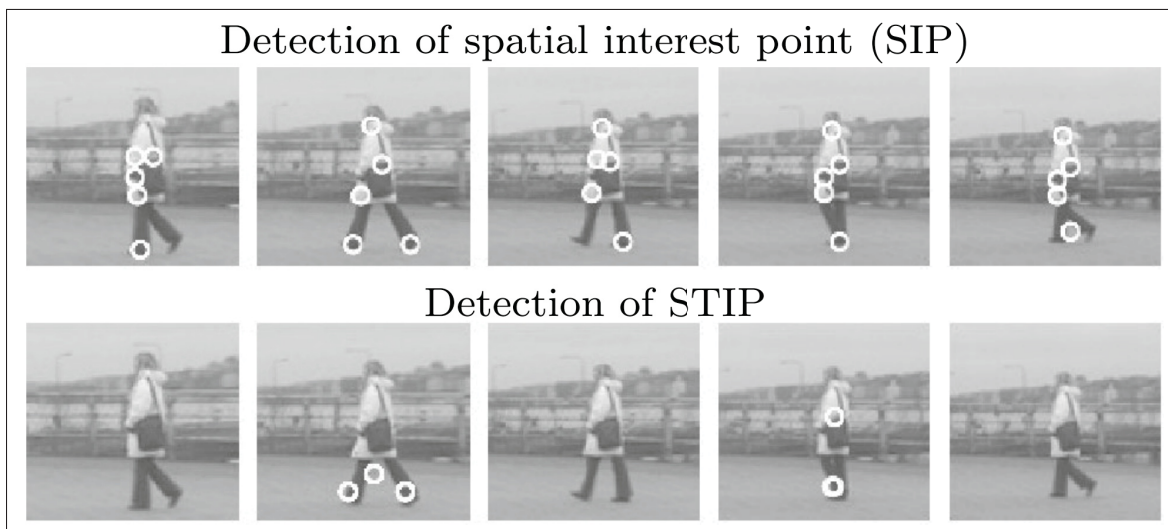


Figure 1.8 Comparaison visuelle sur les points d'intérêt spatiaux SIP avec STIP
(Das Dawn & Shaikh, 2016)

Les descripteurs Histogram of Oriented Gradients (HOG) et Histogram of Optical Flow (HOF) ont été adaptés au domaine temporel pour caractériser les patterns de mouvement. HOG capture

l'information de forme statique dans chaque frame, tandis que HOF encode les patterns de mouvement via le flot optique. La combinaison de ces descripteurs dans une représentation unifiée permet de capturer à la fois l'apparence et la dynamique des actions. L'approche Bag of Words, héritée du traitement du langage naturel, a été largement appliquée à la reconnaissance d'actions. Cette méthode construit un vocabulaire visuel en clustering les descripteurs locaux extraits des vidéos d'entraînement. Chaque action est ensuite représentée par un histogramme d'occurrences de ces "mots visuels", permettant l'application de classificateurs standards comme les machines à vecteurs supports. Cependant, ces approches manuelles présentaient des limitations importantes. Leur dépendance à des descripteurs prédéfinis limitait leur capacité d'adaptation à de nouveaux domaines ou types d'actions. De plus, la conception de descripteurs efficaces nécessitait une expertise domaine considérable et demeurait souvent sous-optimale pour des tâches spécifiques.

1.4.3 Architectures modernes et innovations récentes

Les architectures contemporaines de reconnaissance d'actions exploitent des innovations architecturales sophistiquées pour améliorer l'efficacité et la précision. I3D (Inflated 3D ConvNets), (Wang, Miao, Zhang & Hao, 2019a), propose une méthode élégante pour initialiser les CNN 3D en "gonflant" les poids de modèles 2D pré-entraînés. Cette approche permet de bénéficier des représentations apprises sur ImageNet tout en étendant leur capacité au domaine temporel. La Figure 1.9 présente la structure du modèle I3D (Inflated 3D ConvNet) illustrant le principe de l'inflation des filtres convolutifs 2D en filtres 3D pour la modélisation spatio-temporelle comme (a) montre le module Inception original avant inflation où les convolutions opèrent uniquement sur les dimensions spatiales (hauteur et largeur) et (b) illustre la version après inflation, dans laquelle les noyaux sont étendus à la troisième dimension temporelle permettant au réseau de capturer simultanément les variations spatiales et dynamiques dans les séquences vidéo. La Figure 1.10 illustre la comparaison des performances du modèle SlowFast par rapport à l'architecture Slow-only sur l'ensemble de données AVA. Le modèle SlowFast atteint une amélioration moyenne de +5,2 mAP, avec des gains particulièrement marqués pour

certaines classes d'actions (+27,4, +18,8, +15,9), démontrant sa capacité à équilibrer précision et vitesse d'analyse.

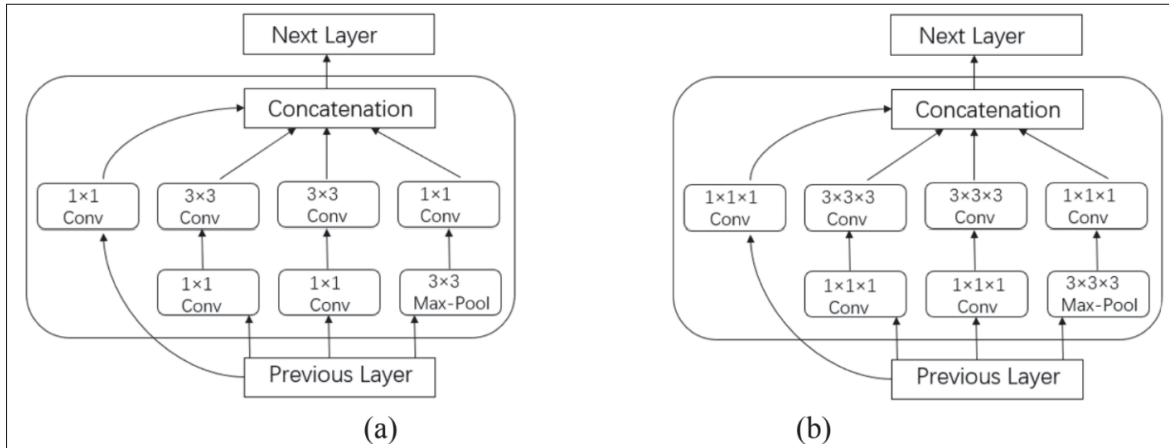


Figure 1.9 I3D (a) module Inception avant inflation, (b) module Inception après inflation
(Wang *et al.*, 2019a)

SlowFast Networks (Feichtenhofer *et al.*, 2019) introduisent une architecture inspirée par les propriétés du système visuel humain. Le pathway "Slow" traite les frames à faible fréquence temporelle pour capturer l'information sémantique et spatiale. Le pathway "Fast" analyse des séquences haute fréquence pour détecter les mouvements rapides. Cette séparation permet d'optimiser chaque flux pour ses caractéristiques temporelles spécifiques. Les Video Transformers adaptent l'architecture Transformer, révolutionnaire en traitement du langage naturel, au domaine vidéo. Ces modèles segmentent les vidéos en patches spatio-temporels traités comme des tokens, permettant l'application des mécanismes d'attention. L'attention temporelle permet de capturer des dépendances à long terme entre frames distants, particulièrement utile pour des actions de longue durée ou présentant des structures temporelles complexes.

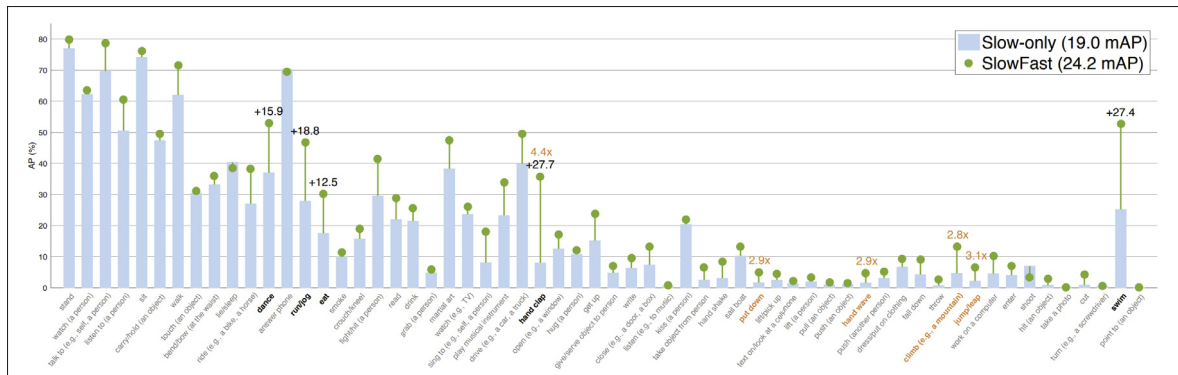


Figure 1.10 AP par catégorie sur AVA
(Feichtenhofer *et al.*, 2019)

1.4.4 Gestion des séquences temporelles et modélisation dynamique

La modélisation efficace de la dimension temporelle constitue un défi central en reconnaissance d'actions. Les Recurrent Neural Networks (RNN), et particulièrement les Long Short-Term Memory (LSTM) networks, ont été largement utilisés pour capturer les dépendances temporelles. Ces architectures maintiennent un état interne qui évolue à travers la séquence, permettant de mémoriser des informations pertinentes sur des durées variables. Les Gated Recurrent Units (GRU) proposent une simplification des LSTM tout en conservant leurs capacités de modélisation temporelle. Cette architecture réduit le nombre de paramètres et accélère l'entraînement, la rendant attractive pour des applications temps réel. Les approches basées sur les graphes offrent une perspective alternative pour modéliser les actions humaines. Spatial-Temporal Graph Convolutional Networks (ST-GCN) (Zhao & Chen, 2023) représentent le squelette humain comme un graphe dont les nœuds correspondent aux articulations et les arêtes aux connexions anatomiques comme présenté dans la Figure 1.11 Cette représentation permet d'appliquer des convolutions graphiques qui respectent la structure naturelle du corps humain tout en capturant l'évolution temporelle des poses.

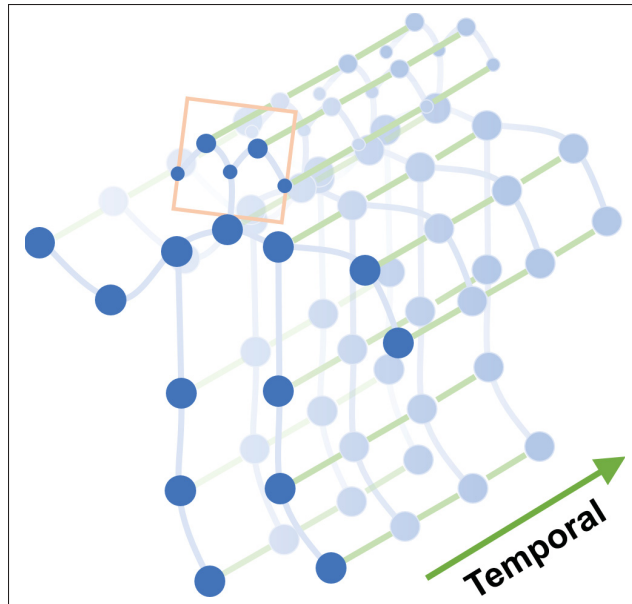


Figure 1.11 Le graphe spatio-temporel d’une séquence squelettique sur lequel opère le ST-GCN proposé (Zhao & Chen, 2023)

1.4.5 Modèle de référence : ProtoGCN

La classification des actions de boxe s’appuie sur l’architecture ProtoGCN qui introduit des innovations importantes pour la reconnaissance d’actions basée sur des poses squelettiques. Cette architecture exploite les Graph Convolutional Networks (GCN) pour modéliser explicitement la structure anatomique du corps humain tout en intégrant un mécanisme de prototypage adaptatif qui améliore la généralisation. L’architecture ProtoGCN représente les séquences de poses comme des graphes spatio-temporels où les nœuds correspondent aux articulations et les arêtes encodent à la fois la connectivité anatomique et les relations temporelles. Cette représentation préserve la structure naturelle des données de mouvement humain et permet aux convolutions graphiques de capturer efficacement les patterns de coordination inter-articulaire caractéristiques de chaque action. Le mécanisme de prototypage constitue l’innovation de ProtoGCN (voir Figure 1.12). Cette approche apprend des prototypes représentatifs pour chaque classe d’action qui capturent les variations intra-classe tout en maintenant la distinctivité inter-classe. Les prototypes

s'adaptent dynamiquement pendant l'entraînement pour optimiser la couverture de l'espace des variations légitimes de chaque technique de boxe. La configuration spécifique à la boxe utilise un graphe à 17 articulations standard COCO. Les arêtes anatomiques connectent les articulations selon la structure squelettique naturelle, tandis que les arêtes temporelles relient chaque nœud à lui-même dans les frames adjacentes. Cette topologie capture efficacement les dépendances spatio-temporelles essentielles à l'analyse des mouvements sportifs.

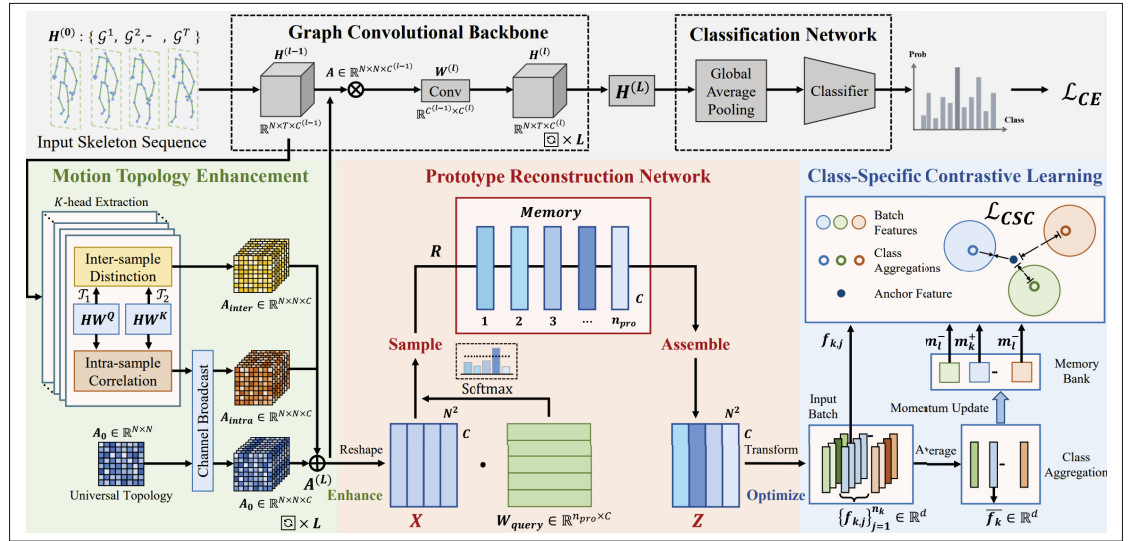


Figure 1.12 Architecture globale de ProtoGCN : Un réseau de reconstruction de prototypes est proposé pour transformer la représentation de la topologie du graphe X en une combinaison Z de prototypes apprenables au niveau des articulations fines (Liu *et al.*, 2025)

1.5 Génération des données synthétiques

Dans le cadre de l'analyse de la boxe, cette stratégie permet de créer des ensembles de données riches, équilibrés et annotés, essentiels à la robustesse des modèles de détection et de reconnaissance d'actions. Cette section présente les motivations de cette approche, les principes de modélisation du corps humain via SMPL, les techniques de capture et de retargeting du mouvement, ainsi que leurs applications spécifiques à la boxe pour la détection de coups et l'analyse tactique de la performance.

1.5.1 Motivations et Enjeux

La génération de données synthétiques émerge comme une solution fondamentale aux limitations inhérentes à la collecte de données réelles dans le domaine de l'analyse du mouvement humain. Cette approche consiste à créer artificiellement des données d'entraînement par simulation numérique, permettant de contourner les contraintes pratiques, éthiques et économiques associées à l'acquisition de données authentiques. Dans le contexte sportif, et particulièrement en boxe, plusieurs facteurs motivent le recours aux données synthétiques. Premièrement, l'accès aux athlètes de haut niveau demeure limité en raison de leurs emplois du temps contraints et des considérations de confidentialité entourant leurs méthodes d'entraînement. Deuxièmement, la variabilité des conditions de capture (éclairage, angles de vue, qualité d'équipement) complique la constitution des jeux de données homogènes et exploitables. Troisièmement, certaines techniques ou situations de combat sont naturellement rares, créant des déséquilibres de classes problématiques pour l'entraînement de modèles d'apprentissage automatique. La génération synthétique offre un contrôle total sur les conditions de création des données. Il devient possible de varier systématiquement les paramètres d'intérêt : morphologies des athlètes, styles de combat, conditions d'éclairage, et points de vue caméra. Cette maîtrise permet de créer des jeux de données équilibrés où chaque classe d'action est représentée de manière homogène, condition essentielle pour l'entraînement de classificateurs robustes.

1.5.2 Modélisation du corps humain : SMPL et ses extensions

Le modèle SMPL constitue une représentation statistique paramétrique du corps humain largement adoptée pour la génération de données synthétiques (Varol *et al.*, 2017). Ce modèle combine un maillage de surface déformable avec des paramètres de forme et de pose, permettant de générer des corps humains diversifiés dans une variété de configurations.

L'utilisation de SMPL comme présenté dans la Figure 1.13 pour la génération de données de boxe offre plusieurs avantages : contrôle précis de la morphologie des combattants, génération automatique d'annotations de pose, et capacité à créer des variations anatomiques représentatives de la diversité des athlètes. Ce modèle paramétrique représente le corps humain via un maillage

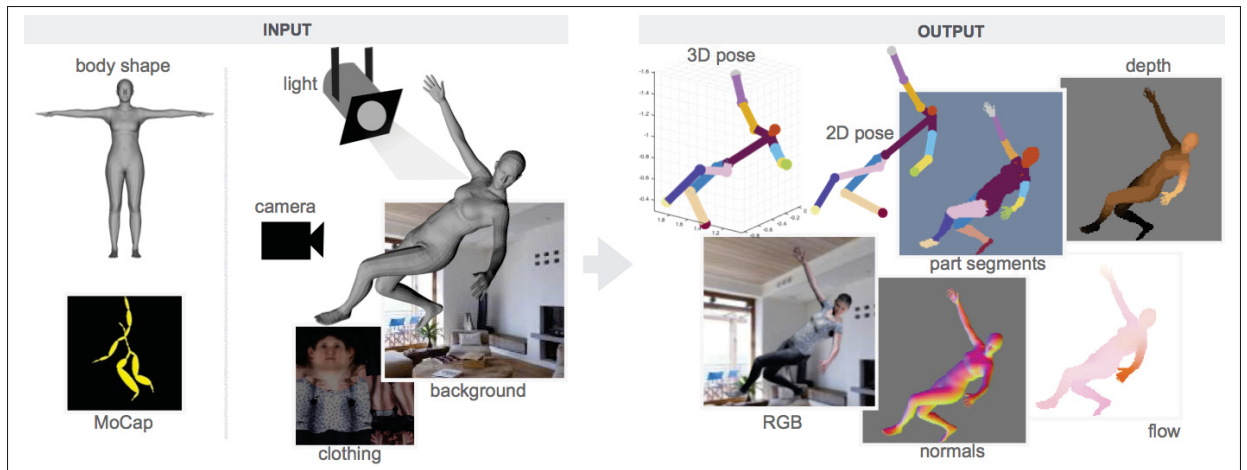


Figure 1.13 Génération d'un jeu de données synthétique à partir de modèles SMPL et de séquences de motion capture (Varol *et al.*, 2017)

triangulaire déformable contrôlé par un ensemble de paramètres de forme et de pose. SMPL modélise la variabilité morphologique humaine à travers un espace de paramètres de forme β de dimension réduite (typiquement 10 dimensions) appris par analyse en composantes principales sur un large corpus de scans corporels. Ces paramètres permettent de générer une grande diversité de morphologies, des individus élancés aux corpulences plus robustes, en respectant les contraintes anatomiques naturelles. La pose du modèle est contrôlée par un vecteur de paramètres θ correspondant aux rotations des 23 articulations du squelette. Ces rotations, exprimées en représentation axis-angle ou quaternions, définissent complètement la configuration spatiale du corps. Le modèle intègre également des correctifs de pose qui ajustent localement la géométrie du maillage pour compenser les déformations non linéaires observées lors de mouvements extrêmes comme la Figure 1.14 illustre le fonctionnement du modèle HMR (Human Mesh Recovery) qui combine apprentissage supervisé et adversarial pour la reconstruction 3D de corps humains à partir d'images monoculaires.

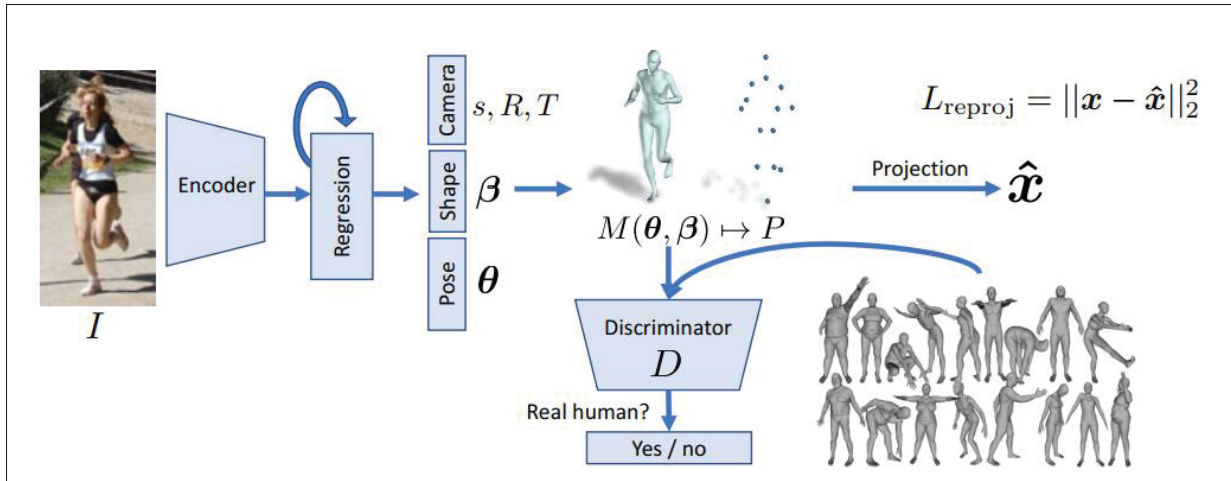


Figure 1.14 Modèle HMR effectue un apprentissage par confrontation sur des données maillées humaines réelles à grande échelle (Varol *et al.*, 2017)

1.5.3 Systèmes de capture de mouvement et retargeting

La génération de mouvements réalistes s'appuie généralement sur des données de capture de mouvement acquises sur des humains. Ces systèmes utilisent diverses technologies : marqueurs optiques passifs ou actifs, capteurs inertiels ou techniques de vision par ordinateur, pour enregistrer les trajectoires articulaires avec une précision temporelle et spatiale élevée. Le processus de retargeting constitue une étape cruciale qui consiste à adapter les mouvements capturés sur un performer spécifique à des modèles virtuels de morphologies différentes. Cette adaptation doit préserver la qualité perceptuelle du mouvement tout en respectant les contraintes anatomiques du modèle cible. Les techniques de retargeting modernes utilisent des approches basées sur la cinématique inverse, optimisant les paramètres articulaires pour reproduire fidèlement des mouvements observés. Des méthodes sophistiquées intègrent des contraintes biomécaniques pour éviter les configurations anatomiquement impossibles et maintenir la plausibilité physique des mouvements générés.

1.5.4 Randomisation et adaptation de domaine

Dans le contexte de la vision par ordinateur, la disparité entre des données synthétiques issues de simulation ou de rendu 3D et des données réelles constitue un obstacle à la généralisation des modèles. Pour atténuer cet écart, deux paradigmes complémentaires se distinguent dans la littérature : la randomisation de domaine (DR) et l'adaptation de domaine (DA) (Tobin *et al.*, 2017; Wang, Li, Ye, Long & Wang, 2019b). Le principe de la randomisation de domaine consiste à générer des données synthétiques en faisant varier de façon aléatoire un large ensemble de paramètres d'environnement et de rendu, tels que les textures, les éclairages, les arrière-plans, les positions de la caméra et les distractions visuelles, afin d'exposer le modèle à la diversité visuelle (Prakash *et al.*, 2019), comme dans la figure 1.15.



Figure 1.15 Exemple de randomisation structurée par domaine (SDR)
(Prakash *et al.*, 2019)

L'hypothèse sous-jacente est que, si le réseau voit suffisamment de variations, la distribution des images réelles sera perçue comme une simple variation parmi d'autres, et le modèle apprendra des caractéristiques invariantes, robustes aux perturbations du monde réel. Cette approche, introduite notamment dans la littérature sur la détection d'objets et l'estimation de pose, a

montré qu'un entraînement exclusif sur données synthétiques randomisées peut suffire à obtenir une bonne généralisation sur des images réelles. D'autres travaux complètent la DR par des mécanismes d'adaptation de domaine visant à aligner les distributions de caractéristiques entre le domaine source (synthétique) et le domaine cible (réel) (Wang *et al.*, 2019b). Ces méthodes (DA) peuvent être supervisées, semi-supervisées ou non supervisées, et reposent souvent sur des techniques d'alignement statistique ou adversarial pour minimiser le décalage entre les représentations extraites des deux domaines. L'approche que nous développons dans cette thèse peut être vue comme une forme de randomisation structurée. En effet, notre procédure de génération synthétique intègre des variations contrôlées mais diversifiées d'éclairage, d'angles de caméra, de positionnements des boxeurs, ainsi que de textures et d'arrière-plans. Ainsi, les vidéos synthétiques visent une diversité suffisante pour contraindre le modèle à apprendre des caractéristiques d'action robustes. L'intérêt de formaliser notre corpus comme DR est multiple. Premièrement, cela nous inscrit dans un cadre méthodologique éprouvé pour réduire la sensibilité aux variations visuelles non pertinentes. Deuxièmement, cela explique pourquoi notre pipeline, entraîné sur des données synthétiques variées réalisées via notre générateur, peut se généraliser raisonnablement sur nos données de boxe. En résumé, notre génération de vidéos synthétiques, par sa diversité contrôlée d'éclairage, d'angles, de positionnements et d'interactions, ainsi que par sa flexibilité, peut être considérée comme une application concrète de la randomisation de domaine.

1.5.5 Applications spécifiques à la boxe

Cette section illustre deux volets clés de ces applications : la détection automatique de coups, qui vise à reconnaître les gestes offensifs et défensifs en conditions réelles ou simulées, et l'analyse tactique, qui s'intéresse à l'interprétation stratégique des séquences de combat pour soutenir la prise de décision et l'entraînement de haut niveau.

1.5.5.1 Détection automatique de coups

Les recherches récentes en détection automatique de coups de poing montrent des résultats prometteurs. (Stefański, Kozak & Jach, 2024) ont développé un système de détection utilisant une seule caméra statique, démontrant la faisabilité de l'analyse automatisée en conditions réelles de compétition. La Figure 1.16 présente le diagramme méthodologique du processus complet de détection automatique de coups de poing illustrant les principales étapes du pipeline, depuis l'enregistrement des vidéos jusqu'à l'évaluation des performances de classification.

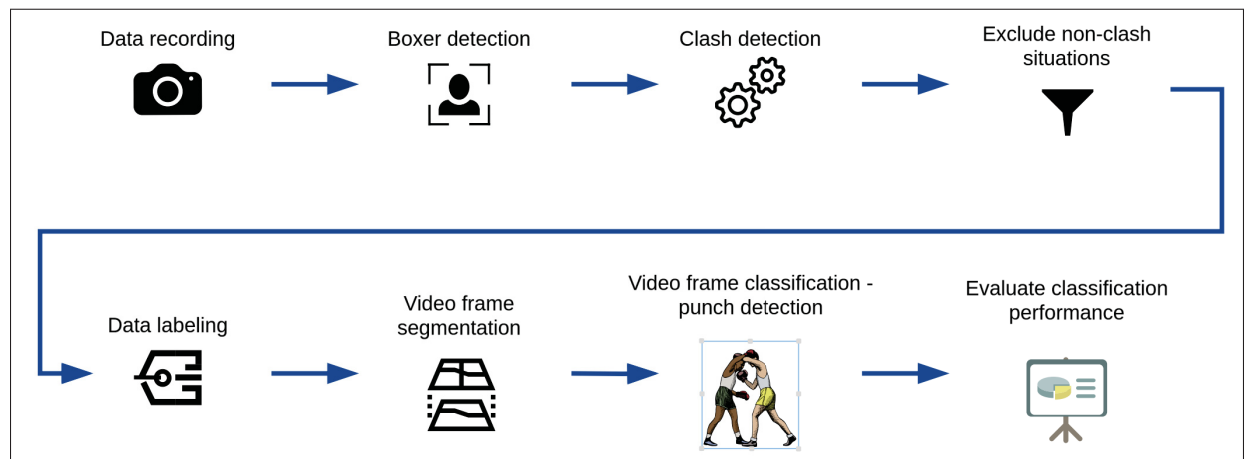


Figure 1.16 Diagramme méthodologique de l'ensemble du processus de traitement (Stefański *et al.*, 2024)

(Kasiri, Fookes, Sridharan & Morgan, 2017) ont proposé un cadre robuste pour la reconnaissance fine des coups de boxe à partir d'images de profondeur aériennes grâce à une stratégie de classification hiérarchique sémantique. La Figure 1.17 présente les résultats de précision obtenus par différentes méthodes de classification appliquées à la détection des coups de poing. Le tableau compare plusieurs variantes de classificateurs SVM celui de one-vs-rest (1vR), one-vs-one (1v1), DAG-SVM et SVM sémantique ainsi qu'un modèle Random Forest évalués sur divers sous-ensembles de caractéristiques (FD, FT, FF, FM, FA, SC, SM). Les résultats montrent que les modèles basés sur SVM surpassent systématiquement Random Forest pour l'ensemble des

jeux de caractéristiques, atteignant des précisions maximales comprises entre 97.0% et 97.3% selon la configuration.

Feature Subset	SVM (1vR)	Random Forest	SVM (1v1)	DAG-SVM	SVM (Semantic)
FD	80.5%	82.8%	89.4%	89.4	89.4%
FT	89.7%	88.4%	93.7%	93.7	94.0%
FF	90.7%	86.1%	94.3%	94.3	94.3%
FM	92.4%	89.1%	95.4%	95.4	95.7%
FA	92.7%	92.1%	97.0%	97.0%	97.0%
SC	92.1%	89.4%	97.3%	97.3%	97.3%
SM	93.4%	89.1%	97.0%	97.0%	97.0%

Figure 1.17 Précision de la classification des coups de poing
(Kasiri *et al.*, 2017)

1.5.5.2 Analyse de performance tactique

L'extension de la reconnaissance d'actions vers l'analyse tactique représente un horizon de recherche prometteur. L'identification automatique des patterns offensifs et défensifs, l'analyse des enchaînements techniques et l'évaluation de l'efficacité comparative des stratégies constituent des applications à forte valeur ajoutée pour l'entraînement sportif, plus de détails sont ajoutés dans l'Annexe III. Dans ce contexte, les travaux de (Seminaro, 2022) ont mis en lumière l'importance d'une approche quantitative de la performance tactique en boxe. L'auteure a proposé une méthode d'extraction automatique de métriques de performance basée sur la vision par ordinateur, capable de classifier les actions offensives avec une exactitude moyenne pondérée de 77 % et de suivre la position des boxeurs avec une précision moyenne de 12,2 cm sur l'aire de combat (Seminaro, 2022). Ces indicateurs tels que le nombre de coups portés, leur taux de réussite et l'occupation spatiale du ring visent à mesurer la domination stratégique

d'un athlète selon les critères de l'Association Internationale de Boxe (IBA). Cependant, les résultats de (Seminaro, 2022) révèlent également plusieurs limites : certaines techniques proches visuellement, comme le crochet arrière ou l'uppercut avant, étaient souvent confondues avec la position de garde, entraînant des erreurs de classification notables. Ces difficultés soulignent la complexité d'une reconnaissance d'actions fiable en conditions réelles et l'importance d'intégrer des données plus variées et des modèles plus robustes pour la généralisation inter-athlètes.

1.6 Synthèse et positionnement de la recherche

Cette revue de littérature met en évidence la convergence progressive entre la génération de données synthétiques, l'estimation de pose et la reconnaissance d'actions vers des solutions intégrées d'analyse sportive automatisée. Ces trois axes ne constituent pas de simples techniques indépendantes, mais des composantes interdépendantes formant la base de notre méthodologie. La génération de données synthétiques offre un moyen concret de pallier la rareté des jeux de données spécialisés en sports de combat et le déséquilibre des classes d'actions. Elle permet de créer un environnement expérimental contrôlé, où les variables visuelles (morphologie, éclairage, caméra, texture) peuvent être systématiquement modulées. Ce choix s'avère essentiel pour notre pipeline, car il garantit une diversité suffisante pour entraîner des modèles robustes aux variations du monde réel tout en réduisant la dépendance à la collecte coûteuse de données annotées. L'estimation de pose 2D, quant à elle, joue un rôle pivot dans la traduction des observations visuelles en informations biomécaniques exploitables. Elle fournit une représentation normalisée du mouvement humain, indispensable pour découpler la dynamique corporelle des conditions visuelles. Dans notre cadre, cette étape constitue le lien entre les images simulées et les séquences réelles, assurant la cohérence entre domaines et la comparabilité des actions. Enfin, la reconnaissance d'actions spatio-temporelles constitue la composante décisionnelle du système. Les architectures modernes comme ST-GCN et ProtoGCN permettent d'exploiter la structure du squelette et les corrélations temporelles pour identifier des actions spécifiques à la boxe. Le choix de ces modèles découle directement de leur capacité démontrée à capturer les relations inter-articulaires et à généraliser sur des mouvements rapides et complexes, typiques des combats.

Ainsi, notre positionnement méthodologique repose sur une intégration cohérente de ces trois piliers : la synthèse de données pour la diversité et la scalabilité, l'estimation de pose pour la structuration biomécanique, et la reconnaissance d'actions pour l'interprétation stratégique. Cette articulation permet de combler les principales lacunes identifiées dans l'état de l'art et de proposer une approche complète, reproductible et adaptée à l'analyse de performance en boxe. de performance sportive.

CHAPITRE 2

MÉTHODOLOGIE ET PIPELINE DE GÉNÉRATION DE DONNÉES SYNTHÉTIQUES

2.1 Introduction

La génération de données synthétiques pour l'analyse de performance en boxe nécessite une approche méthodologique rigoureuse qui intègre de multiples composantes technologiques. Notre pipeline, développé en collaboration avec l'Institut national du sport du Québec, vise à créer un environnement de simulation complet capable de générer des séquences vidéo réalistes de combats de boxe avec des annotations précises. L'architecture générale de notre système s'articule autour de quatre modules principaux interconnectés. Le premier module concerne la modélisation et l'animation des boxeurs virtuels, exploitant le modèle SMPL pour créer des représentations anatomiquement cohérentes. Le deuxième module gère la capture et le retargeting des mouvements de boxe, transformant les données de motion capture en animations exploitables. Le troisième module orchestre la simulation des environnements de combat, incluant la modélisation des rings, l'éclairage et les conditions de caméra. Enfin, le quatrième module assure le rendu final et la génération automatique des vidéos synthétiques annotés.

2.2 Architecture générale du système

Cette section décrit l'architecture générale du système développé pour la génération et l'analyse des données de boxe. L'organisation du flux de données entre ses différents composants et la gestion des paramètres permettant d'assurer la flexibilité et la reproductibilité des expériences (Voir Annexe IV). La Figure 2.1 illustre notre pipeline général conçu pour la génération et l'analyse des données synthétiques de boxe développées dans le cadre de ce projet. Ce schéma met en évidence la structure modulaire du système, depuis la création des environnements 3D jusqu'à l'évaluation finale des performances du modèle d'analyse. Le processus débute par l'intégration des éléments principaux comme les boxeurs et le ring de boxe dans un environnement virtuel. Ces entités sont ensuite soumises à une phase de randomisation des paramètres visuels et contextuels (postures, éclairage, textures, angles de caméra) visant à accroître la diversité du

jeu de données. La phase suivante, le rendu (rendering), permet la génération des séquences vidéo synthétiques. Les données produites sont ensuite exploitées dans la phase d'estimation de pose et de suivi d'identités (ID Tracking) afin de garantir la continuité des trajectoires des boxeurs à travers les trames. Ces informations alimentent un classificateur d'actions, entraîné sur les annotations manuelles et automatiques issues de la phase d'annotation du jeu de données. Enfin, les résultats sont évalués au moyen d'un module d'analyse de performance permettant de mesurer la précision de la reconnaissance d'actions. Ainsi, ce pipeline offre une approche complète et pour la génération de données synthétiques et leur exploitation dans des scénarios d'apprentissage profond appliqués à l'analyse des combats de boxe.

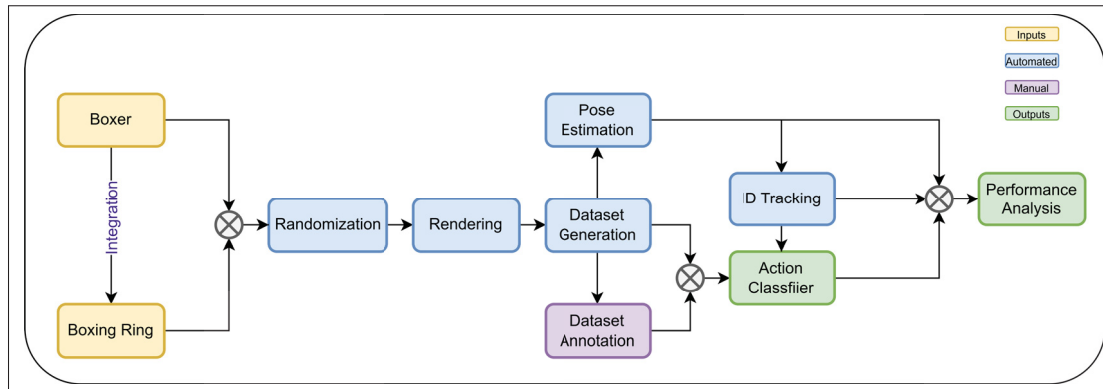


Figure 2.1 Le pipeline général de la génération des données synthétiques

2.2.1 Conception modulaire et flux de données

L'architecture du système repose sur une conception modulaire visant à faciliter l'adaptation et l'évolution du pipeline. Le flux de données suit une séquence cohérente, allant des entrées brutes jusqu'aux sorties finales utilisées pour l'analyse. Le module de gestion des assets constitue le point d'entrée du système. Il regroupe l'ensemble des ressources nécessaires à la génération : modèles 3D des boxeurs, textures corporelles et vestimentaires, fichiers de motion capture, paramètres d'environnement et configurations de rendu. Cette organisation permet une gestion structurée des versions et garantit la cohérence entre les différents jeux de données. Le moteur de simulation est le module central. Il combine les données provenant des autres composants

pour produire les scènes 3D et contrôler la temporalité des animations. Il intègre également les algorithmes de randomisation qui permettent de varier les conditions de rendu, les positions et les interactions entre éléments, afin d’augmenter la diversité des données produites. Enfin, le pipeline de rendu convertit les scènes 3D en séquences d’images 2D. Il applique les modèles d’éclairage, les paramètres de caméra et les traitements visuels requis pour obtenir les rendus.

2.2.2 Gestion des paramètres

La complexité du système nécessite une gestion adaptée des paramètres de configuration. Nous avons développé un système hiérarchique qui organise les paramètres selon leur niveau d’abstraction et leur fréquence de modification. Cette organisation facilite l’expérimentation et permet une adaptation rapide à différents contextes d’utilisation. Les paramètres globaux définissent les caractéristiques générales de génération : résolution des vidéos après le rendu, le nombre de caméras utilisées et les politiques de randomisation. Ces paramètres restent généralement constants au session de génération donnée. Les paramètres de scène contrôlent les aspects spécifiques à chaque séquence générée comme la configuration du ring, conditions d’éclairage, positions des caméras et caractéristiques des boxeurs. Ces paramètres varient selon les besoins de diversification du jeu de données.

2.2.2.1 Styles de rings de boxe

Dans le contexte de la boxe, le style et la conception de l’arène ou de l’environnement du ring peuvent être influencés par divers facteurs tels que le type de l’événement, son importance culturelle ou encore certaines réglementations spécifiques. Voici des exemples de différents styles que nous avons intégrés dans nos scènes de boxe dans le cadre de la création de notre jeu de données synthétiques. La Figure 2.2 présente le premier aperçu général de ring de boxe intégrés dans les scènes synthétiques générées. L’exemple présenté dans la figure illustre notamment un ring institutionnel inspiré du design de l’INS Québec utilisé dans les phases de simulation et de rendu 3D. Ce type d’environnement contribue à la crédibilité visuelle des scènes synthétiques et à la généralisation du modèle d’apprentissage lors de son application sur des vidéos de combats.

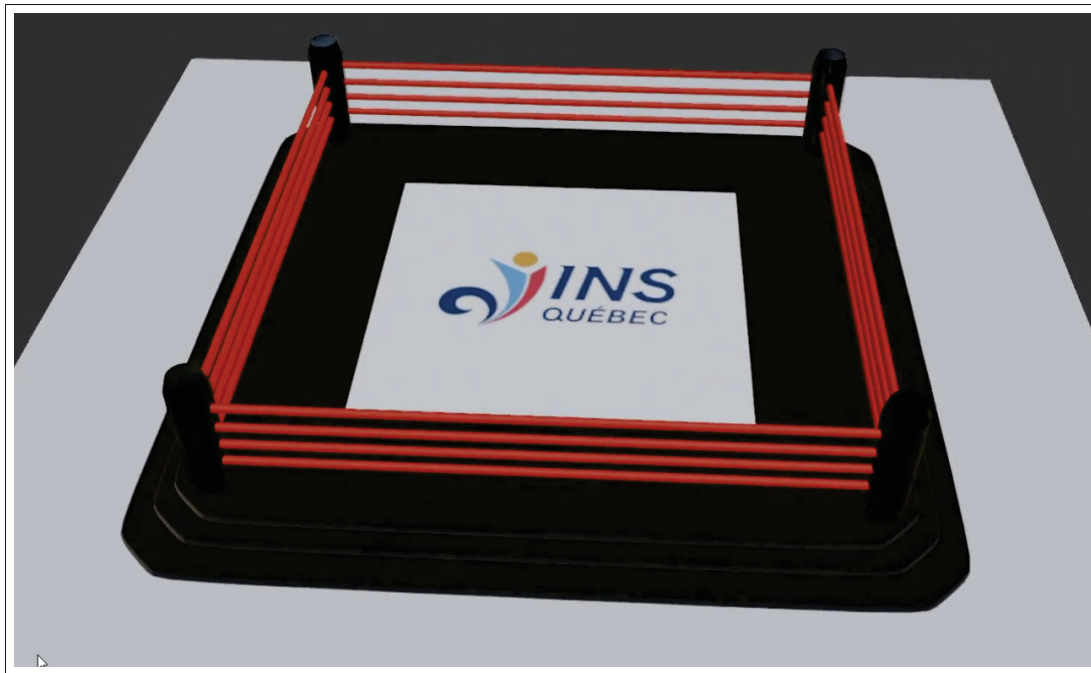


Figure 2.2 Aperçu général de rings de boxe illustrant la diversité des environnements intégrés

Ring de boxe style Las Vegas. Les rings de Las Vegas comme dans la Figure 2.3 sont souvent associés à des combats professionnels de haut niveau et à de grands événements en pay-per-view. Ils se distinguent par :

- **Éclairage :** projecteurs puissants et théâtraux dirigés sur le ring, créant un effet dramatique ;
- **Couleurs et matériaux :** cordes brillantes à fort contraste, typiquement rouges, blanches et bleues.



Figure 2.3 Aperçu de ring de boxe dans le style Las Vegas

Ring de boxe style olympique. Les rings utilisés aux Jeux Olympiques comme dans la Figure 2.4 respectent les normes établies par l'Association Internationale de Boxe (IBA/AIBA). Ces rings sont standardisés et conçus pour assurer la sécurité des athlètes. Ils se caractérisent par :

- **Dimensions :** typiquement 6.1 m (soit 20 ft) de côté, avec trois jeux de cordes autour du périmètre ;
- **Éclairage :** éclairage uniforme et puissant pour garantir une visibilité constante ;
- **Couleurs et agencement :** palette de couleurs neutres ou sobres, tapis souvent blanc ou bleu.

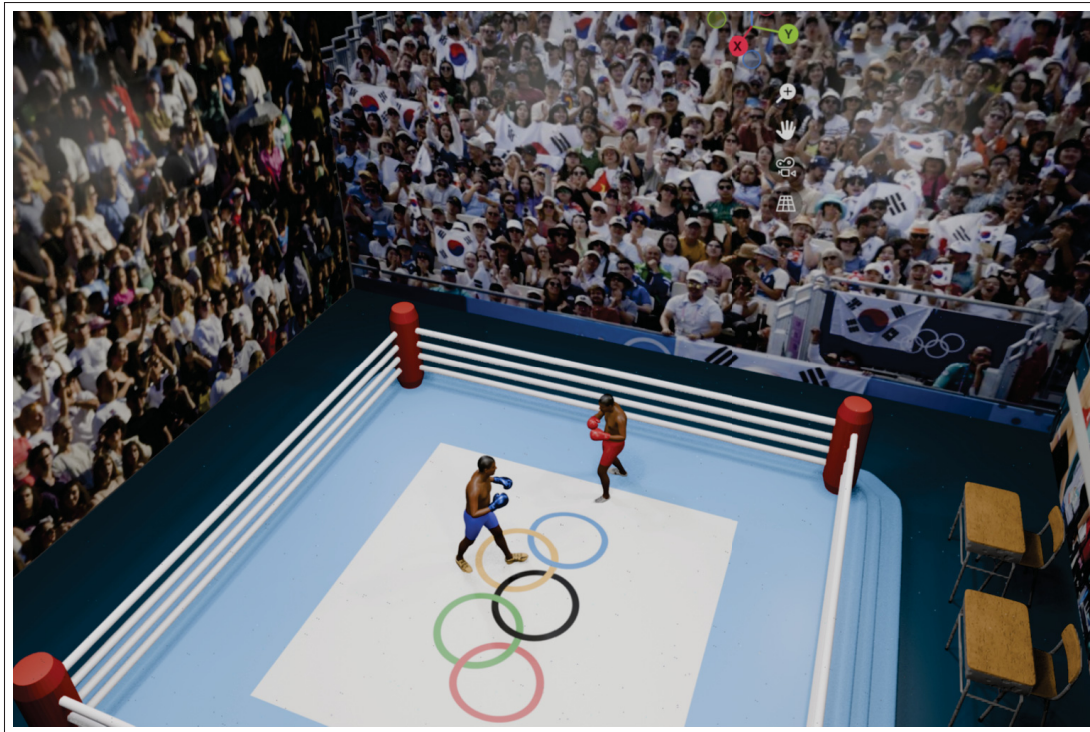


Figure 2.4 Aperçu de ring de boxe dans les Jeux Olympiques

Ring de boxe style salle d'entraînement. Ce style représente les rings des salles de boxe locales comme dans la Figure 2.5 utilisés pour l'entraînement et le sparring plutôt que pour les événements officiels. La priorité est donnée à la fonctionnalité plutôt qu'à l'apparence. Il se caractérise par :

- **Simplicité** : conception minimaliste avec des tapis usés ou unis, cordes standard ;
- **Éclairage** : éclairage fluorescent basique ou lampes suspendues simples ;
- **Environnement** : équipement de gym visible autour du ring (sacs de frappe, poires de vitesse, haltères) ;
- **Couleurs** : tons neutres ou sobres, plus orientés vers la pratique que l'esthétique.



Figure 2.5 Aperçu de ring de boxe dans une salle d'entraînement

La Figure 2.6 illustre une vue d'ensemble du jeu de données synthétiques généré à partir de multiples points de vue de caméra autour du ring. Cette représentation met en évidence la diversité des angles de capture utilisés pour reproduire les conditions d'observation réelles rencontrées lors des compétitions ou des séances d'entraînement. Chaque scène a été rendue à partir d'un ensemble prédéfini de caméras positionnées à différentes hauteurs et orientations sur les côtés du ring afin d'assurer une couverture complète de l'espace de combat. Cette variation des perspectives favorise la robustesse du modèle d'estimation de pose et de reconnaissance d'actions en l'exposant à des contextes visuels variés comme les changements de profondeur, occlusions partielles et perspectives obliques. Ainsi, cette approche multi-caméra permet de constituer un jeu de données riche et équilibré indispensable pour l'entraînement de modèles de vision par ordinateur capables de généraliser efficacement à des configurations réelles de capture vidéo dans le domaine de la boxe.

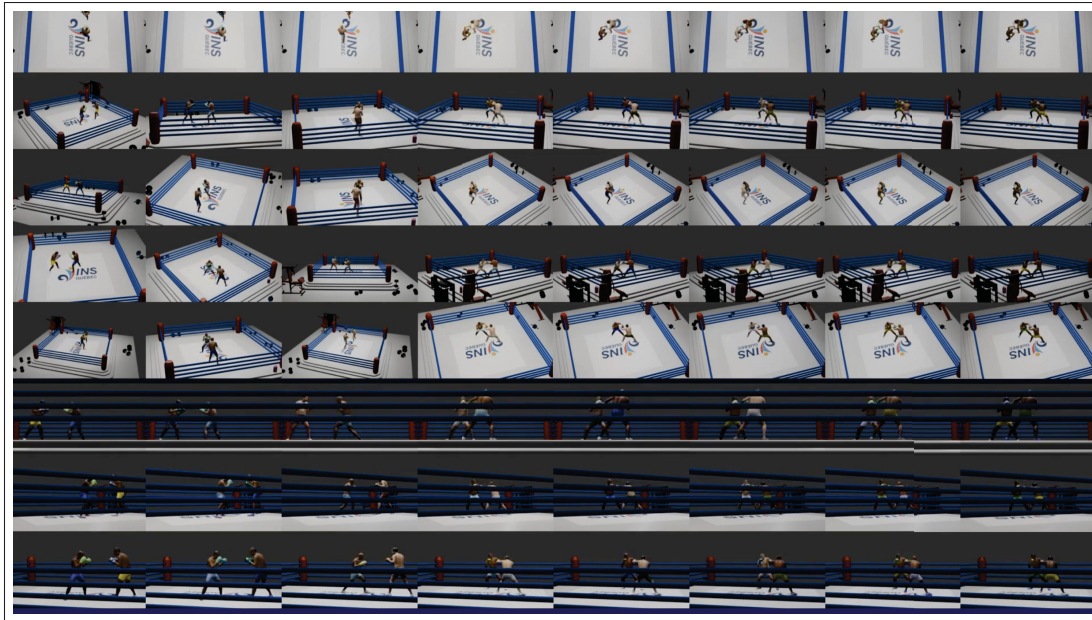


Figure 2.6 Vue d'ensemble du jeu de données synthétiques généré à partir de différents points de vue

2.2.2.2 Caméras et éclairages

Le placement stratégique des caméras virtuelles constitue un aspect crucial pour la génération de données d'entraînement efficaces. Les différents points de vue offrent des informations complémentaires qui enrichissent la capacité des modèles à généraliser vers de nouvelles conditions de capture. Notre approche développe des stratégies de placement qui maximisent la diversité des perspectives tout en préservant la qualité informative des images générées. L'analyse des pratiques de captation vidéo en boxe révèle plusieurs catégories de points de vue standards. Les caméras ringside, positionnées au niveau du ring et orientées horizontalement, fournissent une perspective naturelle qui correspond à la vision des spectateurs et des entraîneurs. Ces vues latérales permettent une analyse fine des techniques en profil et facilitent l'évaluation des distances de combat. Les caméras surélevées, placées en hauteur avec une inclinaison vers le bas, offrent une vue plongeante qui minimise les occultations entre boxeurs. Cette perspective s'avère particulièrement utile pour l'analyse des déplacements et des schémas tactiques, révélant les aspects stratégiques parfois masqués par les vues horizontales.

Étant donné une scène 3D contenant des caméras et des sources lumineuses, l'objectif est d'optimiser leurs positions n, m et leurs propriétés (angles, intensités, distances) de manière à :

- maximiser la couverture visuelle de la scène tout en minimisant les occultations ;
- assurer une distribution d'éclairage homogène réduisant les ombres indésirables.

La formulation théorique repose sur la fonction objectif suivante :

$$O = \alpha \cdot \sum_i C_i(V_i, S) + \beta \cdot \sum_j L_j(P_j, S) - \gamma \cdot \sum_i \sum_j \text{ShadowCost}(V_i, P_j, S) \quad (2.1)$$

où :

- $C_i(V_i, S)$ représente le score de couverture de la caméra i sur la scène S ;
- $L_j(P_j, S)$ quantifie la qualité d'illumination de la lumière j ;
- $\text{ShadowCost}(V_i, P_j, S)$ pénalise les zones d'ombre visibles depuis la caméra i ;
- α, β, γ pondèrent respectivement l'importance de la couverture, de l'éclairage et de la réduction des ombres.

Dans la pratique, la résolution exacte de cette fonction via une optimisation continue serait coûteuse et difficilement intégrable dans un pipeline de rendu. Ainsi, la mise en œuvre a adopté une approche équivalente sur le plan conceptuel mais simplifiée sur le plan computationnel traduisant ce cadre théorique en une méthode de placement paramétrique.

Les caméras sont positionnées autour du centre du ring selon une distribution sphérique variable (variation des azimuts, élévations, distances et focales), garantissant une couverture multi-angle sans recourir à une optimisation explicite. Cette stratégie inspirée du terme $C_i(V_i, S)$ maximise la diversité visuelle tout en limitant les zones d'occlusion.

L'éclairage est défini par une combinaison de sources : une lumière *area* centrale pour l'uniformité globale, des *spotlights* dirigés vers le ring pour renforcer les contrastes, et des lumières ponctuelles positionnées aux coins pour atténuer les ombres. Ces placements reproduisent les effets visés par le terme $L_j(P_j, S)$ tout en réduisant le coût d'ombre ShadowCost par l'emploi de paramètres de diffusion douce (*soft shadows*) et de sources multiples.

Ainsi, le modèle théorique et son implémentation conservent la même logique : maximiser

la visibilité et la cohérence lumineuse tout en minimisant les zones non observables et les contrastes excessifs. Cette approche hybride garantit un équilibre entre rigueur théorique et efficacité pratique pour la génération de scènes synthétiques réalistes.

2.2.3 Intégration avec Blender open source

Notre choix de Blender comme plateforme de développement s'appuie sur plusieurs considérations techniques et stratégiques. Blender offre un écosystème de développement mature avec des APIs Python complètes qui facilitent l'automatisation et la personnalisation. Sa nature open source garantit la pérennité et l'évolutivité de notre solution, tout en permettant une adaptation fine aux besoins spécifiques du projet. L'intégration avec Blender exploite ses capacités avancées de modélisation, d'animation, et de rendu tout en étendant ses fonctionnalités par des scripts personnalisés. Cette approche hybride combine la robustesse d'un logiciel éprouvé avec la flexibilité nécessaire à l'implémentation d'algorithmes spécialisés. Le système de plugins développé encapsule les fonctionnalités spécifiques à la génération de données de boxe dans des modules réutilisables. Ces plugins gèrent la création automatique de scènes, l'application de paramètres de randomisation, et l'export des données d'annotation. Cette modularité facilite la maintenance et permet l'extension du système vers d'autres sports de combat.

2.2.4 Harmonisation des Systèmes de Coordonnées entre Blender et les Données de Capture Réelle

L'intégration des données de capture de mouvement dans l'environnement Blender nécessite une harmonisation rigoureuse des systèmes de coordonnées pour assurer la cohérence spatiale entre les séquences réelles et les modèles virtuels. Cette problématique critique implique la résolution de plusieurs défis techniques fondamentaux, notamment la détermination précise de l'orientation verticale ("up") et de la direction avant ("forward") dans l'espace 3D, conformément aux conventions de Blender qui utilise un système de coordonnées où Z représente l'axe vertical et Y l'axe de profondeur. La gestion des rotations et quaternions suit les principes établis dans la littérature spécialisée, particulièrement les méthodes de conversion entre représentations

axis-angle et matrices de rotation pour préserver l'intégrité des transformations spatiales lors du transfert des données SMPL vers l'environnement de rendu. Notre pipeline de traitement exclut systématiquement le tracking des "end sites" car ces éléments, utilisés uniquement pour la visualisation et la détermination de la longueur des os, introduisent des artefacts dans les calculs de retargeting et peuvent compromettre la précision des animations finales. Cette standardisation des systèmes de coordonnées garantit que les mouvements capturés conservent leur authenticité biomécanique tout en respectant les contraintes techniques de l'environnement de développement Blender, permettant ainsi une intégration transparente des données de capture dans notre pipeline de génération de séquences synthétiques.

2.3 Modélisation et animation des boxeurs virtuels

La modélisation et l'animation des boxeurs virtuels constituent une étape centrale de notre méthodologie, visant à reproduire fidèlement la gestuelle, la dynamique et les conditions visuelles d'un combat de boxe. Cette phase combine des approches de modélisation paramétrique, de retargeting de mouvements et de simulation d'environnements physiques afin de construire un pipeline complet et automatisé de génération de données synthétiques. Cette section détaille les étapes successives de ce processus, depuis l'intégration du modèle SMPL jusqu'aux stratégies d'optimisation computationnelle du rendu.

2.3.1 Intégration du modèle SMPL

L'intégration du modèle SMPL dans notre pipeline de génération nécessite une adaptation pour répondre aux spécificités de l'analyse de boxe. Le modèle SMPL standard, bien qu'anatomiquement précis, nécessite des extensions pour capturer fidèlement les caractéristiques des athlètes de combat. La paramétrisation du modèle SMPL s'appuie sur deux types de variables principales. Les paramètres de forme β , au nombre de dix, contrôlent la morphologie générale du corps en encodant les variations principales observées dans la population humaine. Ces paramètres permettent de générer une diversité morphologique réaliste, des athlètes élancés aux gabarits plus robustes, en respectant les corrélations anatomiques naturelles. Les paramètres de pose θ ,

correspondant aux rotations des vingt-trois articulations principales, définissent la configuration spatiale instantanée du corps. Ces rotations, exprimées en représentation axis-angle pour éviter les singularités des angles d'Euler, permettent de reproduire fidèlement les postures complexes observées en boxe. Notre implémentation intègre des contraintes articulaires physiologiques pour éviter les configurations anatomiquement impossibles. La forme finale est calculée par :

$$V_{\text{shaped}} = \bar{V} + \sum_{i=1}^{10} \beta_i \cdot S_i \quad (2.2)$$

où $\bar{V}$ représente le maillage moyen, β_i les coefficients de forme, et S_i les composantes principales de déformation de forme. Chaque composante principale S_i est une matrice de dimension 6890×3 représentant les déplacements des sommets (x, y, z) par rapport au maillage moyen (Loper *et al.*, 2015). Ces paramètres permettent de générer une diversité morphologique réaliste, des athlètes élancés aux gabarits plus robustes, en respectant les corrélations anatomiques naturelles.

Les paramètres de pose θ , correspondant aux rotations des vingt-trois articulations principales ($24 \times 3 = 72$ paramètres), définissent la configuration spatiale instantanée du corps (Chen *et al.*, 2023). Ces rotations s'expriment mathématiquement par :

$$\theta = [\theta_1, \theta_2, \dots, \theta_{24}] \in \mathbb{R}^{72} \quad (2.3)$$

avec chaque $\theta_i \in \mathbb{R}^3$ représentant une rotation en représentation axis-angle. La conversion axis-angle vers matrice de rotation utilise la formule de Rodrigues :

$$R = I + \sin(\|\theta\|) K + (1 - \cos(\|\theta\|)) K^2 \quad (2.4)$$

où K est la matrice antisymétrique associée au vecteur $\theta/\|\theta\|$, et $\|\theta\|$ l'angle de rotation. Cette représentation évite les singularités des angles d'Euler et permet de reproduire fidèlement les postures complexes observées en boxe.

L'équation complète du modèle SMPL s'exprime par :

$$M(\beta, \theta) = W \left(T_P \left(V_{\text{shaped}}, \theta \right), J(V_{\text{shaped}}), \theta, W \right) \quad (2.5)$$

où :

- T_P applique les déformations dépendantes de la pose :

$$V_{\text{posed}} = V_{\text{shaped}} + \sum_{j=1}^{23} P_j (R_j - I) \quad (2.6)$$

- $J(V_{\text{shaped}})$ calcule les positions articulaires par régression linéaire,
- W réalise le Linear Blend Skinning final.

Les déformations dépendantes de la pose s'expriment par :

$$P(\theta) = \sum_{j=1}^{23} (R_j(\theta) - I) p_j \quad (2.7)$$

où $R_j(\theta)$ est la matrice de rotation de l'articulation j , I la matrice identité, et p_j les blend shapes correctifs spécifiques à chaque articulation.

Notre implémentation intègre des contraintes articulaires physiologiques pour éviter les configurations anatomiquement impossibles (contraintes de vitesse) :

$$\theta_{\min} \leq \theta \leq \theta_{\max} \quad \text{et} \quad \|\nabla \theta\| \leq v_{\max} \quad (2.8)$$

2.3.2 Habillage et apparence des boxeurs virtuels

Une fois le modèle SMPL intégré et paramétré, l'étape suivante consiste à lui conférer une apparence réaliste par un processus de skinning et d'habillage. Pour ce faire, nous avons appliqué sur chaque corps SMPL une peau réaliste et une chevelure issue du jeu de données

BEDLAM (Black *et al.*, 2023), qui fournit des textures haute fidélité associées à des mouvements corporels et des détails anatomiques précis. Cette intégration a permis d'obtenir des modèles présentant une continuité visuelle cohérente entre la géométrie corporelle et la texture cutanée, essentielle pour la crédibilité visuelle des athlètes virtuels. Les textures de surface ont ensuite été enrichies par un processus de coloration et de modification des pixels afin de simuler différents vêtements de boxe. Les shorts, débardeurs et autres éléments vestimentaires ont été créés à partir des cartes de texture du corps, en ajustant les teintes et les zones colorées directement sur les UV du modèle SMPL. Ce procédé permet de générer rapidement une grande diversité d'équipements sans compromettre la cohérence morphologique des boxeurs. Les gants et les chaussures de boxe, quant à eux, ont été importés depuis des modèles open source disponibles en ligne. Ces accessoires ont été adaptés et parentés au squelette du modèle SMPL dans Blender 4, assurant ainsi leur synchronisation complète avec les mouvements articulaires lors des animations. L'ensemble du pipeline aboutit à des boxeurs virtuels entièrement animés et habillés, combinant rigueur anatomique, réalisme visuel et cohérence cinématique, indispensables pour la génération de scènes d'entraînement et de compétition exploitables en analyse biomécanique et en apprentissage profond. Chaque image illustre la combinaison des textures corporelles (peau et cheveux) et des pixels colorés représentant les vêtements, générés à partir du pipeline de rendu synthétique. Ces textures dans la Figure 2.7 permettent de reproduire visuellement la diversité des apparences humaines des boxeurs.



Figure 2.7 échantillon des textures d'avatars boxeurs

2.3.3 Système de rigging et contrôle d'animation

Le système de *rigging* constitue l'interface fondamentale entre le modèle géométrique *SMPL* et les données d'animation issues de la capture de mouvement. Dans notre implémentation, un squelette de contrôle pour les mouvements de boxe a été conçu en intégrant des mouvements avec des contraintes physiques. Le rig personnalisé repose sur une extension de la hiérarchie standard du modèle *SMPL*, à laquelle sont ajoutés des os de contrôle intermédiaires au niveau des épaules, du bassin et des poignets. Ces éléments supplémentaires permettent un contrôle plus fin des rotations articulaires, une meilleure transmission des forces et une amélioration significative de la

qualité visuelle des déformations lors des impacts et des déplacements rapides. L'implémentation de la *Inverse Kinematics* (IK) sur les chaînes cinématiques principales assure la cohérence des contacts et le respect des contraintes spatiales tout au long des séquences de mouvement. Les chaînes IK des membres supérieurs permettent de maintenir avec précision la position des poings lors des frappes, tandis que celles des membres inférieurs garantissent un ancrage stable et réaliste des appuis au sol. Cette structure hiérarchique hybride, combinant *Forward Kinematics* et *Inverse Kinematics*, associe la précision du contrôle direct des articulations du corps.

2.3.4 Capture et retargeting de mouvements de boxe avec Rokoko

Les données de capture de mouvement sont traitées à l'aide de module complémentaire *Rokoko Studio Live* pour Blender (Rokoko, 2023). Cet outil assure le reciblage (retargeting) automatique des séquences enregistrées vers le modèle paramétrique SMPL, en établissant une correspondance hiérarchique entre le squelette du boxeur réel et celui du personnage virtuel. Le processus transfère les rotations articulaires, les vitesses angulaires et la temporalité des gestes tout en ajustant automatiquement les différences morphologiques entre les deux squelettes. Un module d'ajustement cinématique final garantit la stabilité des postures et le respect des contraintes biomécaniques de base, notamment au niveau des articulations majeures (épaules, coudes, genoux). Cette étape permet d'obtenir des animations cohérentes et physiologiquement plausibles, prêtes à être intégrées dans le pipeline de rendu et d'analyse. Le résultat du retargeting constitue ainsi la base de toutes les séquences synthétiques de boxe préservant la dynamique des mouvements capturés.

2.3.5 Simulation d'environnements de combat

La création d'environnements de combat réalistes nécessite une modélisation géométrique qui capture fidèlement les caractéristiques des installations sportives réelles. Notre approche développe une bibliothèque paramétrique de rings de boxe qui couvre la diversité des configurations rencontrées en pratique, depuis les salles d'entraînement modestes jusqu'aux arènes de compétition. La modélisation géométrique s'appuie sur les standards officiels de boxe

pour définir les dimensions de base. Un ring olympique standard mesure 6,1 mètres de côté avec une surface de combat délimitée par des cordes tendues sur quatre niveaux. Notre modèle paramétrique permet de faire varier ces dimensions dans des plages réalistes pour refléter la diversité des installations existantes. La surface de combat utilise une modélisation physiquement réaliste qui influence directement la dynamique des mouvements, particulièrement lors des déplacements rapides et des changements de direction. Les propriétés de friction et de rebond sont calibrées selon des mesures réalisées sur des rings réels. Les éléments structurels (poteaux, cordes, tablier) intègrent des modèles de matériaux appropriés qui affectent l'éclairage et les reflets de la scène. Les cordes utilisent des shaders spécialisés qui simulent la réflexion spéculaire caractéristique des matériaux synthétiques sous éclairage intense. L'éclairage constitue un facteur critique pour la qualité des données générées. Les conditions d'éclairage influencent directement la performance des algorithmes de vision par ordinateur, particulièrement pour la détection de contours et l'estimation de poses. Notre système développe des modèles d'éclairage adaptatifs qui reproduisent fidèlement la diversité des conditions rencontrées en pratique. La modélisation des sources lumineuses s'appuie sur des mesures photométriques réalisées dans différents types d'installations sportives. Les salles d'entraînement utilisent généralement un éclairage fluorescent diffus qui génère des ombres douces et une répartition relativement homogène. Les arènes de compétition exploitent des projecteurs directionnels haute intensité qui créent des contrastes marqués et des ombres nettes. Notre système implémente un modèle d'éclairage physiquement basé qui simule la propagation de la lumière variée, les réflexions multiples, et les phénomènes d'ombrage. Ce modèle utilise des techniques de ray-tracing pour les interactions lumineuses complexes entre les différents éléments de la scène. La variabilité temporelle de l'éclairage simule les fluctuations observées dans les installations réelles. Les variations de tension électrique, le vieillissement des ampoules et les réflexions sur les surfaces mobiles créent des variations subtiles, mais perceptibles de l'illumination. Notre système intègre des variations illimitées par des modulations calibrées sur des observations réelles.

2.3.6 Génération des boxeurs virtuels

La méthodologie proposée repose sur deux grandes phases complémentaires : la génération de boxeurs virtuels réalistes et la mise en place d'un pipeline complet dédié à l'analyse des performances à partir de vidéos synthétiques et réelles. La première phase consiste à produire des modèles humains virtuels animés en s'appuyant sur le modèle SMPL auquel sont associés des mouvements réalistes extraits de fichiers FBX contenant des séquences d'animation issues de captations réelles. Le processus débute par la fusion SMPL–FBX où le modèle SMPL est animé à partir de fichiers FBX décrivant les mouvements d'un boxeur, eux-mêmes dérivés de fichiers XML provenant de sessions de capture de mouvement. Cette étape permet de transférer la dynamique corporelle observée dans les données sources vers le modèle paramétrique. Les mouvements extraits des fichiers FBX font ensuite l'objet d'un rétargeting automatique sur le modèle SMPL effectué dans le respect de la cinématique articulaire et des contraintes biomécaniques du corps humain. Ce transfert garantit la précision des postures et la fluidité des séquences animées, assurant ainsi une correspondance fidèle avec les mouvements d'origine. À cette étape s'ajoute une phase de texturation personnalisée réalisée à partir de textures corporelles genrées (masculines et féminines) et enrichies par une étape de pixelisation et de coloration. Enfin, les textures personnalisées sont combinées à des éléments d'équipement tels que les gants, les chaussures et les shorts de boxe pour créer des avatars complets. L'ensemble est intégré dans un modèle SMPL animé et habillé prêt à être inséré dans des environnements de simulation 3D. Chaque boxeur virtuel ainsi généré constitue une instance unique animée et cohérente sur le plan visuel pouvant être utilisée pour la génération de jeux de données synthétiques ou l'analyse automatique de performances sportives. La Figure 2.8 illustre le pipeline complet de cette création depuis la capture du mouvement jusqu'à la production finale du boxeur virtuel en passant par les étapes de retargeting, de texturation et d'habillage, soulignant la modularité et la reproductibilité de l'approche proposée.

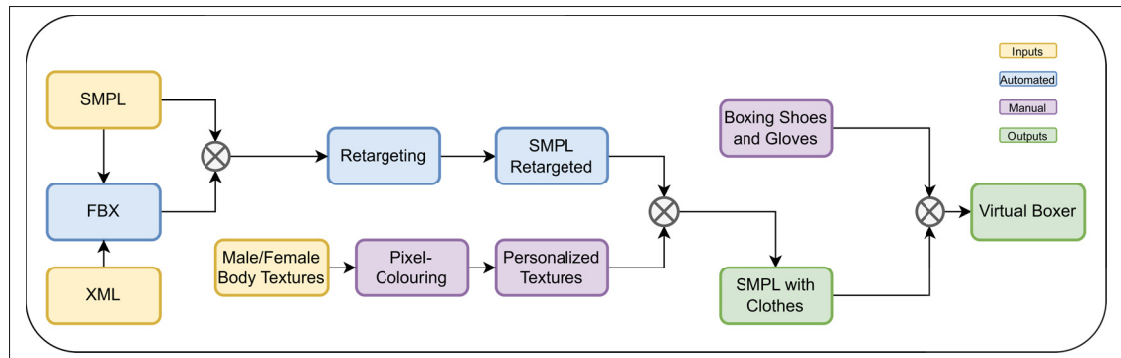


Figure 2.8 Le pipeline de la création du boxeur virtuel

2.3.7 Pipeline de rendu

Le pipeline de rendu constitue l'étape finale de transformation des scènes 3D en images exploitables pour l'entraînement de modèles d'apprentissage automatique. Cette étape vise à produire des rendus cohérents tout en maintenant une efficacité de calcul adaptée à la génération d'un grand volume de données synthétiques. L'implémentation du rendu repose sur le moteur Eevee de Blender qui est un moteur de rendu en temps réel fondé sur la rasterisation. Ce choix de moteur permet de réduire considérablement les temps de calcul tout en conservant un niveau de visuel suffisant pour l'entraînement de réseaux de vision. Les paramètres de rendu sont configurés pour un compromis optimal entre qualité et performance : résolution Full HD (1920×1080), rendu vidéo à 60 fps, éclairages dynamiques et matériaux PBR standardisés. Le modèle de matériau utilisé repose sur la BRDF Principled, qui permet de représenter de manière unifiée les propriétés de surface des différents objets de la scène notamment la peau du boxeur, les textiles des équipements, les surfaces métalliques et le sol du ring. Ces matériaux assurent une cohérence visuelle à travers toutes les scènes générées. Le processus de rendu est automatisé par un script Python intégré dans Blender qui ouvre chaque scène initialise les caméras et lumières, puis effectue le rendu séquentiel des vues. Pour chaque caméra, une séquence vidéo est exportée au format MP4 (codec H.264) garantissant la compatibilité directe avec les pipelines d'entraînement en apprentissage profond. Cette génération multi-vue permet d'obtenir des variations d'éclairage de perspective et de composition visuelle essentielles à la robustesse des

modèles d'analyse d'action. Enfin, des techniques d'optimisation adaptatives ajustent la qualité du rendu selon la complexité géométrique et lumineuse de la scène. EEEVEE exploite la puissance du GPU pour calculer les effets de post-processing (bloom, ambient occlusion, motion blur) tout en maintenant un haut débit de génération des vidéos synthétiques. En moyenne, le rendu complet d'une variation de scène comprenant huit vidéos d'une durée en vion une minute nécessite jusqu'au 1h30 de rendu via l'API du Blender. Ces performances démontrent l'efficacité du pipeline EEEVEE pour la création massive de séquences synthétiques tout en maintenant une cadence de production adaptée aux besoins expérimentaux du projet.

2.3.8 Stratégies d'optimisation computationnelle

La génération de volumes importants de données synthétiques nécessite des optimisations afin de maintenir des temps de traitement raisonnables tout en préservant la qualité et la cohérence des résultats. Les scripts de génération ont été développés et testés initialement sur une station mobile personnelle équipée d'un processeur Intel Core i9-12900HK (12e génération) de 64 Go de mémoire vive et d'une carte graphique NVIDIA GeForce RTX 3050 Ti. L'exécution finale des campagnes de génération a été réalisée sur la station de travail du laboratoire de capture de mouvement (MoCap Lab) de l'ÉTS dotée d'une carte graphique NVIDIA GeForce RTX 4090 et de 128 Go de mémoire vive. Cette infrastructure a permis de confirmer la bonne scalabilité du pipeline avec des gains de performance proportionnels aux capacités matérielles permettant d'obtenir une accélération supplémentaire d'environ 45% lors du rendu. La gestion des tâches parallèles dans Blender 4 repose sur des scripts Python exploitant le module multiprocessing, permettant l'exécution simultanée de plusieurs processus indépendants de rendu. Chaque instance prend en charge la génération d'une séquence synthétique distincte incluant les personnages, les vêtements, les environnements, les conditions lumineuses et les angles de caméra. Cette exécution parallèle, combinée à une file d'attente dynamique, réduit le temps total de génération tout en maintenant la cohérence et la reproductibilité des résultats. La génération de données s'appuie également sur des workflows automatisés conçus pour minimiser les interventions manuelles répétitives. Le système intègre des fonctionnalités de scripting avancées permettant la

création et la gestion de campagnes de génération à grande échelle avec une supervision minimale. Le traitement par lots (batch processing) autorise la définition de campagnes composées de multiples configurations exécutées de manière séquentielle.

2.4 Génération des annotations de vidéos

Dans le cadre de l'analyse de la boxe, cette phase d'annotation garantit la cohérence entre les données visuelles et les catégories d'actions définies par les experts du domaine (Voir Annexe II). Cette section présente l'outil d'annotation développé spécifiquement pour ce projet, ainsi que la méthodologie appliquée à l'annotation des vidéos de boxe, en mettant l'accent sur la fiabilité, la reproductibilité et la conformité aux standards de recherche en vision par ordinateur.

2.4.1 Notre outil d'annotation

Notre outil d'annotation vidéo a été spécifiquement développé pour analyser et annoter les mouvements de boxe dans le cadre de notre recherche sur la génération de séquences synthétiques. Il s'agit d'une application web interactive, construite avec des technologies modernes (Node.js, Vite), permettant une annotation précise et structurée des données de capture de mouvement.

L'interface propose une vue multi-caméras simultanée avec quatre angles de prise de vue différents de la même scène de boxe, offrant une couverture complète de l'espace d'entraînement selon notre protocole de capture utilisant quatre caméras disposées en configuration cylindrique. Cette interface est complétée par une timeline permettant une navigation temporelle précise avec des frames numérotées, des contrôles de lecture, ainsi que des indicateurs visuels pour marquer les segments annotés.

Le système d'annotation structuré comprend plusieurs niveaux détaillés (voir Annexe I) :

- **Actions techniques :** Jab (coup direct rapide du poing avant), Cross (coup direct puissant du poing arrière), Lead Hook (crochet circulaire du poing avant), Rear Hook (crochet circulaire du poing arrière), Lead Uppercut (uppercut remontant du poing avant), Rear Uppercut (uppercut remontant du poing arrière), Rear Overhand (coup descendant du poing arrière).

- **Mouvements défensifs** : Slip (esquive latérale de la tête), Bob and Weave (esquive par flexion et déplacement), Pull (retrait rapide vers l'arrière), Duck (esquive par accroupissement), Parry (déviation défensive), Block (blocage direct), Clinch (corps à corps défensif).
- **Déplacements tactiques** : Pivot (rotation sur un pied), Angle off (sortie angulaire).
- **Feintes** : Punch Feint (simulation de coup), Shoulder Feint (feinte d'épaule), Feet/Step Feint (feinte de pas).
- **Positions** : Neutral (garde de base).

Les résultats d'actions sont catégorisés selon leur efficacité (par exemple coups portés au corps, bloqués, esquivés, ou dans le vide), ainsi que par les postures corporelles critiques pour l'analyse biomécanique :

- **Balance** : désigne une posture dans laquelle le poids du corps est bien réparti, soit de manière neutre entre les deux jambes, soit volontairement déplacé vers l'avant ou l'arrière pour s'adapter à une action imminente (attaque ou défense). Cette position est essentielle pour maintenir une bonne stabilité dynamique, optimiser les appuis au sol, et permettre une transition fluide entre offensive et esquive. Elle reflète également l'état de préparation du boxeur à générer de la puissance ou à se déplacer efficacement.
- **Low hands** : position où les bras sont tenus plus bas que la garde traditionnelle. Cette posture expose davantage la tête du boxeur, mais peut servir à inciter l'adversaire à attaquer ou à améliorer la mobilité des bras pour des attaques rapides.
- **Southpaw stance** : posture adoptée par les boxeurs gauchers, avec le bras droit en avant et le bras gauche en arrière. Cette configuration inverse la dynamique habituelle des échanges et oblige l'adversaire à s'adapter à des angles inhabituels.
- **Orthodox stance** : posture standard utilisée par les boxeurs droitiers, avec le bras gauche en avant pour le jab et le bras droit en arrière pour les coups puissants. Elle permet un bon équilibre entre attaque et défense.
- **Squared** : posture où le torse du boxeur est orienté de face vers l'adversaire, avec les pieds presque parallèles. Bien que cette position permette une plus grande surface de frappe et des

échanges directs, elle augmente également l'exposition aux attaques adverses et réduit la mobilité latérale.

et un système d'identification individuelle des boxeurs (*person0*, *person1*) avec attribution spécifique des actions, le tout structuré en fichiers JSON intégrables dans notre pipeline.

Cette approche d'annotation semi-automatisée suit un workflow complet :

1. segmentation temporelle (définition des frames de début et de fin),
2. classification hiérarchique (attribution des labels action → outcome → posture → personne),
3. export structuré.

Elle constitue une étape cruciale pour :

- la validation de cohérence biomécanique,
- la constitution de jeux de données annotés pour l'entraînement des modèles,
- l'analyse documentée des standards techniques et des variations individuelles

Ainsi, notre outil permet de créer des jeux de données annotés de référence de haute qualité pour valider les performances de nos algorithmes de classification et d'optimisation. L'évolution de notre classification initiale comportant vingt actions distinctes vers une taxonomie plus concise définie par le tableau Tableau-A-I-1 illustre une stratégie à la fois scientifique et pragmatique en reconnaissance d'actions. Cette transition repose sur les fondements biomécaniques du geste pugilistique et répond à des exigences d'implémentation en contexte réel, en particulier dans une optique de collaboration avec les méthodologies établies par l'INS Québec.

Hiérarchie des actions offensives : La classification des gestes de frappe repose sur une hiérarchie biomécanique. Les coups directs comme le *jab* (catégorie 1) et le *cross* (catégorie 2) sont essentiels, utilisés à la fois pour l'attaque et la défense, avec un engagement minimal pour le jab, et une puissance accrue pour le cross. Les crochets (*lead/rear hook*) et uppercuts (*lead/rear uppercut*) se distinguent par des rotations corporelles complètes et une coordination segmentaire spécifique, respectivement. La catégorie générique *punch* regroupe les coups puissants moins spécifiques, tels que les overhands ou certaines feintes.

Catégories défensives : Les actions défensives sont séparées en deux grandes classes. D’une part, les actions de blocage (*block*), représentant une part significative des gestes défensifs observés dans notre base de données. D’autre part, les techniques d’évitement (*evade*) comme le *bob and weave*, les esquives latérales et le *duck* traduisent une stratégie de mobilité active pour éviter les coups.

Modèles de déplacement tactique : Le contrôle de la position dans l’espace est représenté par la classe *move*, qui inclut les déplacements, les changements de garde et le *clinch*. Ces mouvements structurent la dynamique du combat. La classe *neutral*, quant à elle, sert de repère de stabilité, définissant les phases de transition entre techniques actives.

Principes issus des sciences du sport : La distinction biomécanique entre les types de coups est bien établie : direction de l’avant-bras, angle du coude, et translation du centre de masse sont des indicateurs déterminants. Les jabs émergent dans des contextes spatiaux différents des uppercuts et crochets, qui mobilisent davantage la rotation du tronc. Une organisation hiérarchique du système de classification est ainsi privilégiée, allant des catégories générales aux actions spécialisées.

Bénéfices de la simplification des classes d’actions de boxe Cette simplification présente plusieurs avantages notables. D’abord, elle assure une cohérence biomécanique en regroupant les actions selon leurs similarités mécaniques et anatomiques, facilitant ainsi l’analyse des mouvements. Ensuite, elle simplifie le traitement informatique, rendant le système plus léger et adapté à une reconnaissance rapide en temps réel. De plus, cette classification reste en phase avec la pratique sportive, offrant aux entraîneurs une taxonomie intuitive et directement exploitable pour l’évaluation technique. Enfin, sa structure hiérarchique garantit une excellente scalabilité, permettant l’ajout progressif de sous-classes ou de niveaux de détail au fur et à mesure de l’évolution du modèle.

Ainsi, il constitue un compromis optimal entre granularité descriptive, rigueur scientifique et efficacité opérationnelle, facilitant l’analyse automatisée des performances en boxe tout en s’adaptant aux exigences du terrain et des experts du domaine.

2.4.2 Annotation des vidéos de boxe

Interface multi-caméras

L'outil intègre présenté dans la Figure 2.9 une visualisation multi-caméras simultanée, présentant quatre angles de vue synchronisés, capturés par des caméras disposées autour du ring. Cette vue multiperspective assure une couverture complète de la zone d'entraînement et facilite l'annotation des actions, postures et mouvements en limitant les ambiguïtés dues aux occlusions ou aux angles morts. La timeline intégrée, accompagnée d'indicateurs visuels et de numéros de frame explicites, permet une navigation fluide et un repérage précis des segments à annoter. Chaque séquence est annotée sur plusieurs niveaux taxonomiques : *action* → *outcome* → *posture* → *personne*, afin de permettre des analyses fines et hiérarchiques.

Processus d'annotation

Le pipeline mis en place repose sur plusieurs étapes reproductibles :

1. **Chargement des vidéos** : les fichiers sources sont importés dans l'outil via le navigateur, selon une organisation par dossier.
2. **Étiquetage Figure 2.10** : les annotateurs attribuent les labels d'action, de posture, de résultat et d'identification du boxeur (*person0*, *person1*) en utilisant les fonctions *add* et *add & advance*, assistées de raccourcis clavier pour optimiser la vitesse de saisie.
3. **Sauvegarde et export** : les annotations sont sauvegardées au format JSON à chaque session (*Ctrl+S*). Chaque fichier est versionné et archivé avec un identifiant unique et un horodatage pour garantir la traçabilité.

L'ensemble du workflow inclut un plan de sauvegarde structuré, et la validation manuelle et automatisée des labels par des heuristiques de cohérence temporelle et spatiale.



Figure 2.9 Interface de l'outil d'annotation avec vue multi-caméras et timeline synchronisée

Contrôles qualité

Enfin, le processus comprend une validation a posteriori des annotations, intégrant :

- la vérification automatique de la cohérence temporelle (alignement timestamps-actions)
- l'application de règles logiques pour identifier les erreurs systématiques
- des scripts d'audit et de correction en lot
- la possibilité d'intégrer des approches semi-supervisées (par exemple label-smoothing) pour prendre en compte l'incertitude lors de l'apprentissage.

L'ensemble de cette méthodologie s'appuie sur les fichiers de consignes de L'INS Québec et garantit une annotation structurée, documentée, traçable et conforme aux standards académiques de la recherche en vision par ordinateur et analyse biomécanique.

```

<mocap shape_0="-0.0020681 -0.0003498 0.0001824 0.0002072 0.0000294 0.0000700 0.0000386
0.0000376 -0.0000522 0.0000609" shape_1="0.0163333 0.0024086 0.0014837 0.0037102 0.0002940
-0.0047204 -0.0006060 0.0015662 -0.0001329 0.0000980">
<keyframe key="000000"><key personID="0" poses="0.6575383 -0.6794847 1.0444651 1.4224839
0.6731099 0.6636714 -0.5985612 0.0728066 0.4381323 -0.0921061 -0.1216443 -0.3998058 0.4034639
0.1450608 0.0416063 0.5040773 0.0482640 -0.0934138 0.6286945 -0.0557893 -0.0538826 0.0743190
-0.0330532 -0.0490138 0.3396854 0.0420064 -0.1737057 -0.1663999 -0.1056968 0.3498541
0.0815411 0.1373608 -0.0149099 -0.4156404 0.1238146 0.2573637 -0.1136722 0.0658758 -0.7298275
0.2031325 0.1455481 -0.2153098 -0.1134177 -0.1559528 -0.0478061 -0.0137134 0.4108646
0.0734263 -0.1481406 0.0598755 0.0366438 0.0683154 -0.4659074 -1.1165081 -0.0049248 0.6251599
0.9583769 0.2205355 -1.8934734 0.9437840 0.1557985 1.9405158 -0.6958746 -0.3114806 -0.2293466
0.1009098 -0.2699961 0.2493937 -0.0070287 -0.1202847 -0.1352953 -0.4865405 -0.1079182
0.1637747 0.3660476" action="Jab" outcome="Punch landed on torso (body shot)" posture="Balance
" /></keyframe>
<keyframe key="000001"><key personID="0" poses="0.6536042 -0.6819129 1.0392036 1.4249761
0.6583291 0.6697517 -0.6129510 0.0692653 0.4368377 -0.1084236 -0.1267622 -0.4177268 0.4179111
0.1465125 0.0443671 0.5402858 0.0441043 -0.0994866 0.6504264 -0.0569908 -0.0555826 0.0789109
-0.0330912 -0.0469778 0.3322906 0.0439717 -0.1737686 -0.1703859 -0.1102812 0.3484982
0.0850008 0.1401217 -0.0128736 -0.4217528 0.1207777 0.2621166 -0.1266959 0.0677962 -0.7349993
0.2018942 0.1413330 -0.1970325 -0.1141464 -0.1470697 -0.0489278 -0.0138899 0.4100634
0.0748498 -0.1424486 0.0649588 0.0540510 0.0713299 -0.4679295 -1.1011840 -0.0136978 0.6375331
0.9481251 0.2295183 -1.8911422 0.9228896 0.1409946 1.9253407 -0.6788540 -0.3159681 -0.2286846
0.1120818 -0.2702161 0.2489736 -0.0145853 -0.1159448 -0.1349910 -0.4862022 -0.1053874
0.1640698 0.3586462" action="Jab" outcome="Punch landed on torso (body shot)" posture="Balance
" /></keyframe>
<keyframe key="000002"><key personID="0" poses="0.6504371 -0.6850226 1.0311611 1.4228025

```

Figure 2.10 Étiquetage hiérarchique des actions techniques et des segments temporels annotés

2.5 Analyse comparative des jeux de données

L'analyse comparative des jeux de données utilisés dans nos expérimentations met en évidence des différences substantielles en termes de taille, de diversité, de modalités d'annotation et de représentativité des actions de boxe. Tout d'abord, le jeu de données employé dans la thèse de (Seminaro, 2022) se limite à des séquences de *shadow boxing* réalisées par un seul boxeur avec un total d'environ 16 minutes d'images annotées. Les annotations reposent sur des coordonnées 3D des articulations extraites en amont par un estimateur de pose. La distribution des classes reste limitée et peu représentative de la variabilité des situations réelles d'un combat puisque l'absence d'interaction entre plusieurs boxeurs empêche de couvrir des scénarios de défense, de contre-attaque ou d'actions combinées.

En revanche, le jeu de données de base utilisé pour l'apprentissage supervisé est constitué d'environ 250 vidéos réelles correspondant à plus de cinq heures d'enregistrements d'entraînement de boxeurs athlètes de l'INS Québec. À ce jeu de données réel s'ajoute un jeu de données

synthétique de base, composé de 1536 vidéos générées automatiquement depuis des caméras placées autour du ring en direction centre de ring. Ce corpus synthétique de base, constitue l'ensemble synthétique utilisé pour l'entraînement du modèle dans ce travail. Il regroupe les séquences visuelles, leurs métadonnées et l'ensemble des annotations d'actions nécessaires à l'apprentissage supervisé. En parallèle, un volume supplémentaire de plus de 8000 vidéos synthétiques a également été produit, mais il demeure non annoté à ce stade. Ces données non étiquetées n'ont pas été intégrées au processus d'apprentissage et représentent un potentiel significatif pour des travaux futurs, en particulier dans des cadres non supervisés ou auto-supervisés. Contrairement au jeu réel initial, le corpus synthétique exploité dans cette thèse met systématiquement en scène deux boxeurs en interaction, ce qui permet de couvrir un spectre d'actions beaucoup plus large, comprenant vingt classes distinctes allant des coups offensifs aux mouvements défensifs et aux postures. En termes de diversité et de représentativité, la combinaison du réel et du synthétique offre déjà une couverture bien plus complète des situations rencontrées en boxe que les travaux antérieurs. La chaîne de génération automatique étant entièrement opérationnelle et extensible, l'augmentation de ce jeu de données dépend désormais principalement des ressources matérielles et du temps de calcul, plutôt que de limites techniques. L'ensemble renforce ainsi la solidité du cadre expérimental et ouvre la voie à des entraînements à plus grande échelle dans des travaux ultérieurs.

CHAPITRE 3

MODÈLES ET ARCHITECTURES IMPLEMENTÉS

3.1 Introduction

Ce chapitre présente la méthodologie d’entraînement des modèles développés pour l’analyse de la performance en boxe. Nous décrivons d’abord l’architecture des réseaux de neurones profonds utilisés, en précisant les choix de conception et les paramètres d’entraînement adoptés. Nous détaillons ensuite les métriques d’évaluation retenues afin de mesurer la qualité des prédictions et la robustesse des modèles. Enfin, nous exposons les premiers résultats obtenus en classification et en analyse spatio-temporelle des actions de boxe, constituant une base pour les analyses plus approfondies du chapitre suivant.

3.2 Pipeline Général d’Estimation de Pose et de Classification des Actions

Le pipeline proposé pour en Figure 3.1 l’estimation de la pose humaine et la reconnaissance d’actions dans des scènes de boxe repose sur une architecture modulaire articulée en quatre étapes principales : détection des boxeurs, suivi et appariement d’identités, estimation de la pose, et classification des actions. Cette approche permet une exploitation efficace des vidéos synthétiques et réelles annotées pour la modélisation des comportements sportifs.

Détection des boxeurs avec YOLOv8 : La détection initiale des personnes est assurée par le modèle YOLOv8 (Sohan, Sai Ram & Rami Reddy, 2024), qui offre un compromis optimal entre vitesse et précision. Ce détecteur d’objets localise les boxeurs à chaque frame sous forme de *bounding boxes*, même en présence d’occlusions, de postures atypiques ou d’interactions rapprochées. L’efficacité de YOLOv8 dans des scènes dynamiques permet d’initier le pipeline avec une base fiable pour les étapes suivantes.

Estimation de la pose avec ViTPose : À partir des régions détectées, le modèle ViTPose (Xu *et al.*, 2022) est utilisé pour estimer les poses 2D des boxeurs. ViTPose repose sur une architecture Transformer visuelle capable de modéliser les dépendances spatiales longues entre les articulations. Chaque personne détectée est représentée par un ensemble structuré de 17 à 34

points clés représentant les articulations majeures. Cette estimation de pose constitue la base de représentation des mouvements dans la phase suivante.

Suivi d'identité par appariement 3D/2D : Pour garantir la cohérence temporelle du suivi d'identité dans les séquences vidéo, nous implémentons une méthode d'appariement entre les projections 2D des articulations 3D (provenant des armatures dans Blender) et les détections YOLO. À chaque frame, une boîte englobante est projetée à partir des keypoints 3D vers le plan image en utilisant les paramètres de la caméra virtuelle. Cette boîte projetée est ensuite comparée à celles détectées par YOLOv8, en utilisant des métriques d'intersection sur l'aire (IoU) et de proximité spatiale. Le meilleur appariement est conservé pour maintenir une correspondance robuste entre les identités dans le monde 3D et les séquences vidéo. Ce procédé améliore significativement la qualité du suivi multi-individu en présence d'actions rapides et de changements de postures.

Classification des actions avec ProtoGCN : Les séquences temporelles de poses sont ensuite encodées comme des graphes spatiaux-temporels, où les nœuds correspondent aux articulations et les arêtes modélisent à la fois la connectivité anatomique et la continuité temporelle. Le modèle ProtoGCN (Liu *et al.*, 2025) est ensuite appliqué pour classifier ces graphes selon les actions de boxe : frappes (jab, uppercut, hook), défenses (slip, block, duck), mouvements (pivot, feinte), ou encore postures (high guard, low hands, squared). ProtoGCN introduit un mécanisme de prototypage adaptatif qui renforce la généralisation aux styles de boxe variés et aux morphologies différentes, ce qui en fait un choix adapté à l'analyse biomécanique dans des environnements réalistes ou synthétiques.

Intégration multi-sources : Ce pipeline est conçu pour fonctionner sur des vidéos issues de notre pipeline de rendu Blender ou de séquences captées en salle. Il peut être déployé sur des séquences mono- ou multi-vues, permettant des analyses à la fois locales (par vue) et globales (fusion spatio-temporelle). L'ensemble de ce système constitue une solution complète pour l'analyse automatisée des performances dans les sports de combat, fournissant des prédictions d'actions, des postures, et des métriques exploitables par les entraîneurs, les chercheurs en biomécanique, et les systèmes d'IA.

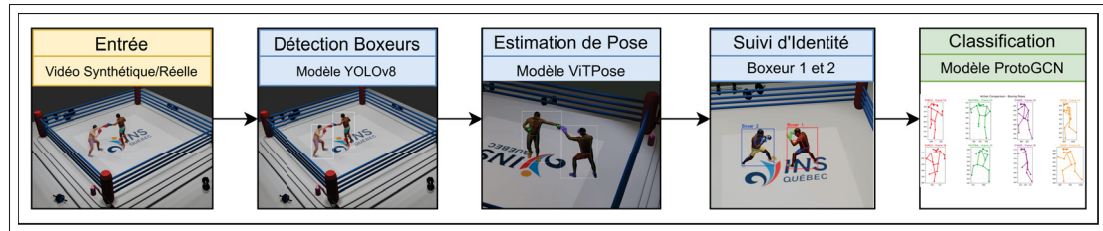


Figure 3.1 Pipeline général d'estimation de pose et de reconnaissance d'action

3.3 Détection et Estimation de la Pose des boxeurs

La détection et le suivi des boxeurs constituent une étape clé du pipeline d'analyse, car ils permettent d'identifier les athlètes dans le ring et de maintenir leur identité d'une image à l'autre. Cette étape est essentielle pour extraire des trajectoires cohérentes et préparer les données nécessaires aux phases suivantes d'estimation de pose et de reconnaissance d'actions.

3.3.1 Architecture ViTPose et Adaptations Spécifiques

L'estimation de pose s'appuie sur l'architecture ViTPose-Large, qui exploite les mécanismes d'attention des Transformers visuels pour modéliser efficacement les dépendances spatiales entre articulations. Cette architecture présente des avantages particuliers pour l'analyse de mouvements sportifs complexes où les relations inter-articulaires distantes jouent un rôle critique dans la précision d'estimation. La configuration d'entrée utilise des images de résolution 384×288 pixels extraites des bounding boxes détectées par YOLOv8. Cette résolution constitue un compromis optimal entre qualité de détail articulaire et efficacité computationnelle. Le prétraitement applique une normalisation standardisée et un padding adaptatif qui préserve les ratios d'aspect naturels tout en assurant l'uniformité des dimensions d'entrée. L'architecture backbone exploite un Vision Transformer (ViT-Large) pré-entraîné sur ImageNet, fournissant une base de représentations visuelles robustes. Le décodeur utilise une architecture pyramidale avec skip connections qui combine efficacement les caractéristiques multi-échelles. Cette conception capture simultanément les détails fins nécessaires à la localisation précise et les patterns globaux

de pose. L'adaptation spécifique à la boxe modifie la tête de prédiction pour optimiser la détection des 17 points clés du format COCO standard.

3.3.2 Méthode d'Appariement 3D/2D

Le suivi d'identité constitue un défi critique dans l'analyse de boxe multi-personnes, particulièrement complexe en raison des interactions rapprochées et des changements rapides de position caractéristiques de ce sport. Notre approche exploite la correspondance entre les modèles 3D de notre simulation Blender et les détections 2D pour maintenir la cohérence temporelle des identités. La méthode d'appariement repose sur la projection géométrique précise des articulations 3D vers le plan image en utilisant les paramètres de caméra calibrés. À chaque frame, les positions articulaires 3D de chaque boxeur virtuel sont transformées en coordonnées 2D par application de la matrice de projection caméra. Cette transformation tient compte des paramètres intrinsèques (focale, centre optique, distorsions) et extrinsèques (position, orientation) pour assurer une correspondance géométrique exacte. L'algorithme d'appariement compare ces projections théoriques avec les détections réelles de YOLOv8 en utilisant une métrique composite qui combine intersection sur union (IoU), distance euclidienne entre centres, et similarité de forme. Cette approche multi-critères améliore la robustesse face aux variations de pose et aux imprécisions de détection. Le score de correspondance intègre ces composantes avec des poids optimisés par validation croisée. L'innovation principale réside dans l'exploitation de l'information 3D pour prédire les positions futures et résoudre les ambiguïtés d'appariement. Lorsque les boxeurs se croisent ou interagissent de près, les trajectoires 3D connues permettent de maintenir les identités correctes même en cas d'occlusions temporaires. Cette prédictibilité constitue un avantage décisif par rapport aux méthodes de suivi purement 2D.

3.3.3 Gestion des Occlusions et Interactions

La gestion des occlusions représente l'aspect le plus critique du suivi en boxe, où les interactions rapprochées créent des masquages complexes qui peuvent perdurer plusieurs secondes. Notre approche développe des stratégies spécialisées qui exploitent à la fois l'information 3D et les

patterns comportementaux typiques de la boxe. La détection d'occlusions utilise une analyse géométrique 3D qui détermine précisément quelles parties de chaque boxeur sont visibles depuis le point de vue caméra. Cette analyse considère non seulement les masquages mutuels entre boxeurs mais aussi les auto-occlusions dues aux rotations corporelles. La stratégie de maintien d'identité pendant les occlusions exploite des prédicteurs de mouvement adaptatifs qui anticipent les trajectoires probables. Ces prédicteurs, calibrés sur les patterns de mouvement observés en boxe, estiment les positions futures avec une précision suffisante pour maintenir l'appariement pendant des occlusions de 2-3 secondes. L'incertitude de prédiction croît graduellement, guidant les décisions de réassociation. L'algorithme de récupération post-occlusion utilise des descripteurs d'apparence robustes qui caractérisent chaque boxeur indépendamment de la pose. Ces descripteurs, basés sur les caractéristiques vestimentaires et morphologiques, permettent la réidentification même après des occlusions prolongées. L'implémentation utilise des réseaux de features appris qui capturent l'identité visuelle distinctive de chaque individu. L'intégration de contraintes temporelles globales optimise l'assignment d'identités sur des fenêtres étendues plutôt que frame par frame. Cette optimisation globale, implémentée par programmation dynamique, résout les conflits d'assignment et minimise les erreurs d'identité sur des séquences complètes.

3.3.4 Détection des personnes dans le jeu de données réel

Afin d'analyser automatiquement les vidéos de boxe capturées en conditions réelles, nous avons mis en place un pipeline de détection et de suivi combinant deux modèles d'apprentissage profond complémentaires. Le modèle RT-DETR (Real-Time Detection Transformer) a été utilisé pour détecter les individus dans chaque trame vidéo. Ce détecteur de type Transformer présente l'avantage d'un traitement direct sans étape de NMS, tout en conservant une rapidité d'inférence adaptée à des flux sportifs. Le modèle ViTPose a ensuite été appliqué aux régions détectées afin d'estimer les points clés articulaires (*keypoints*) des boxeurs. Ce modèle repose sur une architecture Transformer visionnelle qui capture efficacement les dépendances spatiales entre articulations.

Les deux modèles sont exploités conjointement pour produire à la fois les boîtes englobantes (bounding boxes) et les squelettes 2D des athlètes.

3.3.5 Contrainte spatiale : détection de la zone du ring

Les vidéos contiennent parfois des spectateurs ou des entraîneurs visibles en arrière-plan. Il était donc nécessaire de restreindre la détection aux seuls individus situés à l'intérieur du ring. Pour cela, une région d'intérêt centrée a été définie dynamiquement toutes les 30 trames. La proportion de pixels correspondant à la couleur caractéristique des cordes du ring (bleu) dans cette zone a été utilisée pour estimer la confiance dans la détection de la zone de combat. Ce filtrage permet de réduire les fausses détections et de garantir que seules les trajectoires des boxeurs actifs sont prises en compte. Les valeurs de paramètres utilisées dans le pipeline complet résultent d'un compromis entre robustesse et précision présenté dans la Figure 3.2.

- **Seuil de confiance** $\tau = 0.30$. Un seuil trop élevé (par exemple $\tau > 0.5$) risquerait de rejeter des détections de boxeurs partiellement occultés ou en mouvement rapide. À l'inverse, un seuil trop faible augmenterait les faux positifs (spectateurs, arbitres). La valeur de 0.30 est cohérente avec les pratiques usuelles en détection d'objets sur COCO et a montré un bon équilibre entre rappel et précision dans nos données.
- **Aire minimale** $A_{\min} = 1500 \text{ px}^2$. Cette contrainte permet d'éliminer les petites boîtes correspondant à des faux positifs (bras d'arbitre dans le champ, spectateurs éloignés). La valeur a été calibrée empiriquement sur la résolution moyenne de nos vidéos (1280×720), de sorte qu'aucun boxeur en situation normale n'ait une aire inférieure à ce seuil.
- **Distance maximale de normalisation** $D_{\max} = 150$. Cette valeur correspond à une distance typique (en pixels) de déplacement inter-trames d'un boxeur sur des vidéos capturées à 25 fps. Elle sert à normaliser la contribution de la distance euclidienne dans la fonction de coût d'association. Des valeurs plus grandes dilueraient l'effet de la proximité spatiale, tandis que des valeurs plus petites rendraient le suivi instable lors de mouvements rapides.
- **Pondération** $\alpha = 0.3$ (distance) et 0.7 (IoU). L'IoU est priorisé car elle reflète la continuité géométrique des boîtes, ce qui est essentiel lorsque les boxeurs se croisent. Cependant, la

distance entre centres garde un rôle (30%) pour éviter les inversions lors de changements brusques d'échelle (un boxeur s'approchant de la caméra).

- **Fréquence de détection : chaque trame.** La détection est effectuée sur toutes les trames afin de maximiser la précision du suivi et d'assurer une cohérence temporelle des identités. Bien qu'un sous-échantillonnage (`detect_every > 1`) réduise le coût, nous avons retenu la fréquence maximale dans le pipeline officiel pour éviter toute perte de données fines, particulièrement importante pour les mouvements de boxe.
- **Estimateur de pose : ViTPose avec $K = 17$ articulations.** Le choix de la version `vitpose-base-simple` et du format COCO à 17 points répond à deux critères : (i) disponibilité d'un modèle pré-entraîné robuste sur les articulations principales, et (ii) compatibilité avec les structures de GCN couramment utilisées dans la littérature (Yan *et al.*, 2018).

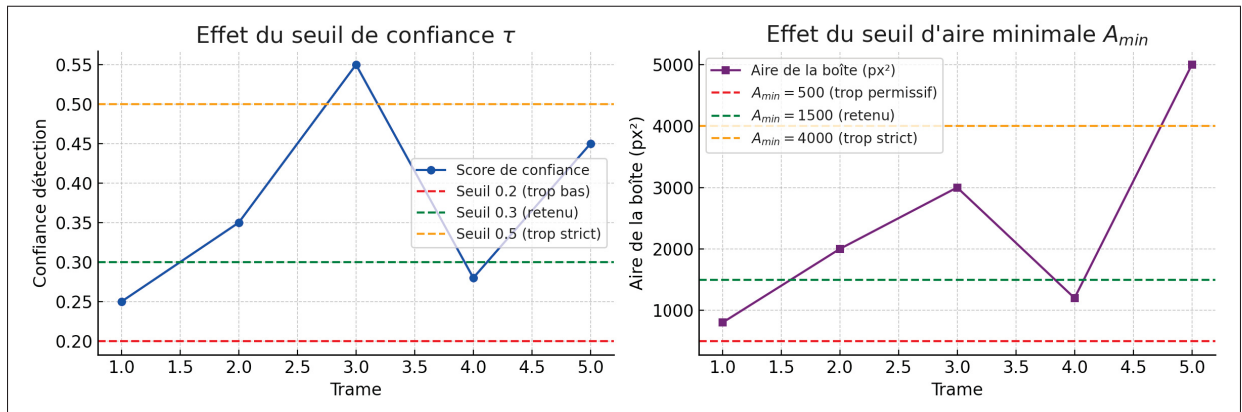


Figure 3.2 Illustration de l'effet des paramètres du pipeline de détection dans les vidéos réelles. Gauche : impact du seuil de confiance τ sur la conservation des détections (un seuil trop bas accepte trop de faux positifs, un seuil trop strict rejette des boxeurs valides).

Droite : effet du seuil d'aire minimale A_{min} sur le filtrage des boîtes (valeurs trop basses conservent des faux positifs, valeurs trop hautes rejettent des boxeurs partiellement visibles)

Ainsi, chaque paramètre du pipeline a été choisi de manière raisonnée, en s'appuyant à la fois sur la littérature, des considérations pratiques liées aux vidéos de boxe, et des calibrations empiriques réalisées sur nos propres données.

3.3.6 Association et suivi des boxeurs

Le suivi des deux boxeurs au fil des trames repose sur une combinaison de métriques géométriques et probabilistes destinées à assurer la continuité temporelle des trajectoires. Deux mesures principales sont exploitées pour évaluer la correspondance entre les détections successives : l'Intersection-over-Union (IoU) et la distance euclidienne entre les centres des boîtes englobantes. La similarité spatiale entre deux boîtes de détection b_1 et b_2 est mesurée à l'aide de l'IoU, définie par :

$$IoU(b_1, b_2) = \frac{|b_1 \cap b_2|}{|b_1 \cup b_2|}, \quad (3.1)$$

où $|b_1 \cap b_2|$ désigne l'aire d'intersection et $|b_1 \cup b_2|$ l'aire d'union entre les deux boîtes. Cette mesure fournit une estimation robuste du recouvrement spatial entre la prédiction courante et l'historique du suivi. Afin de renforcer la cohérence géométrique, la distance entre les centres respectifs $c_1 = (x_1, y_1)$ et $c_2 = (x_2, y_2)$ des boîtes b_1 et b_2 est également considérée :

$$d(c_1, c_2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}, \quad (3.2)$$

où (x_i, y_i) représentent les coordonnées du centre de chaque boîte de détection à l'image.

L'association optimale entre les détections présentes et les trajectoires existantes est ensuite formulée comme un problème d'affectation bipartite résolu à l'aide de l'algorithme de Hongrie. Le coût global de correspondance est défini comme :

$$C = \alpha \cdot \frac{d(c_1, c_2)}{D_{\max}} + (1 - \alpha) (1 - IoU(b_1, b_2)), \quad (3.3)$$

où $\alpha \in [0, 1]$ est un facteur de pondération contrôlant l'équilibre entre distance et recouvrement, et $D_{\max}$ représente la distance maximale de normalisation dans l'image.

Cette formulation garantit une association stable et cohérente même dans des situations complexes comme les croisements ou les occlusions partielles entre les deux boxeurs.

3.3.7 Sorties générées

Pour chaque séquence vidéo, deux types de sorties peuvent être produits. D’une part, une vidéo annotée affichant les boîtes englobantes et les squelettes en surimpression est générée afin de permettre une validation visuelle du bon fonctionnement du pipeline. D’autre part, un fichier CSV structuré rassemble les coordonnées des points clés extraits pour chaque trame, selon le format :

(frame, person_id, kp_index, x, y, score),

où chaque ligne correspond à un point clé identifié, associé à un individu et à une trame donnée. L’ensemble de ces données constitue la base d’entrée du pipeline d’analyse subséquent dédié à la reconnaissance d’actions et à l’évaluation des performances techniques des boxeurs.

3.4 Classification des Actions

La classification des actions constitue l’étape finale du pipeline d’analyse où les séquences de poses estimées sont traduites en catégories sémantiques représentant les différentes techniques de boxe. Cette section présente le modèle de référence utilisé ainsi que les adaptations méthodologiques et architecturales qui ont été introduites pour optimiser sa performance dans le contexte spécifique de la boxe présenté dans la Figure 3.3. L’adaptation de ProtoGCN aux caractéristiques uniques de la boxe nécessite plusieurs modifications architecturales et méthodologiques qui optimisent les performances pour cette application spécialisée. Ces adaptations exploitent les connaissances expertes du domaine pour guider l’apprentissage vers les aspects les plus discriminants. La définition des classes d’actions reflète la taxonomie technique établie en collaboration avec les experts de l’INS Québec. Les huit catégories principales incluent les frappes directes (jab, cross), les frappes circulaires (hook gauche, hook droit), les frappes ascendantes (uppercut gauche, uppercut droit), et les actions défensives (blocage, esquive). Cette classification équilibre la variété technique avec la faisabilité de reconnaissance automatique. L’architecture de convolution graphique utilise trois couches successives avec des dimensions cachées de [64, 128, 256] qui capturent progressivement des patterns de complexité croissante.

La première couche détecte les mouvements articulaires élémentaires. La seconde identifie les coordinations entre segments adjacents. La troisième capture les patterns globaux caractéristiques de chaque technique. Cette hiérarchie suit naturellement l'organisation de l'analyse biomécanique. Le mécanisme d'attention temporelle pondère l'importance des différentes phases du mouvement selon leur valeur discriminative. Cette attention apprend automatiquement à se concentrer sur les phases critiques comme l'armement et l'impact qui caractérisent le mieux chaque technique. L'implémentation utilise des poids d'attention soft qui préservent la différentiabilité tout en permettant une sélectivité temporelle efficace.

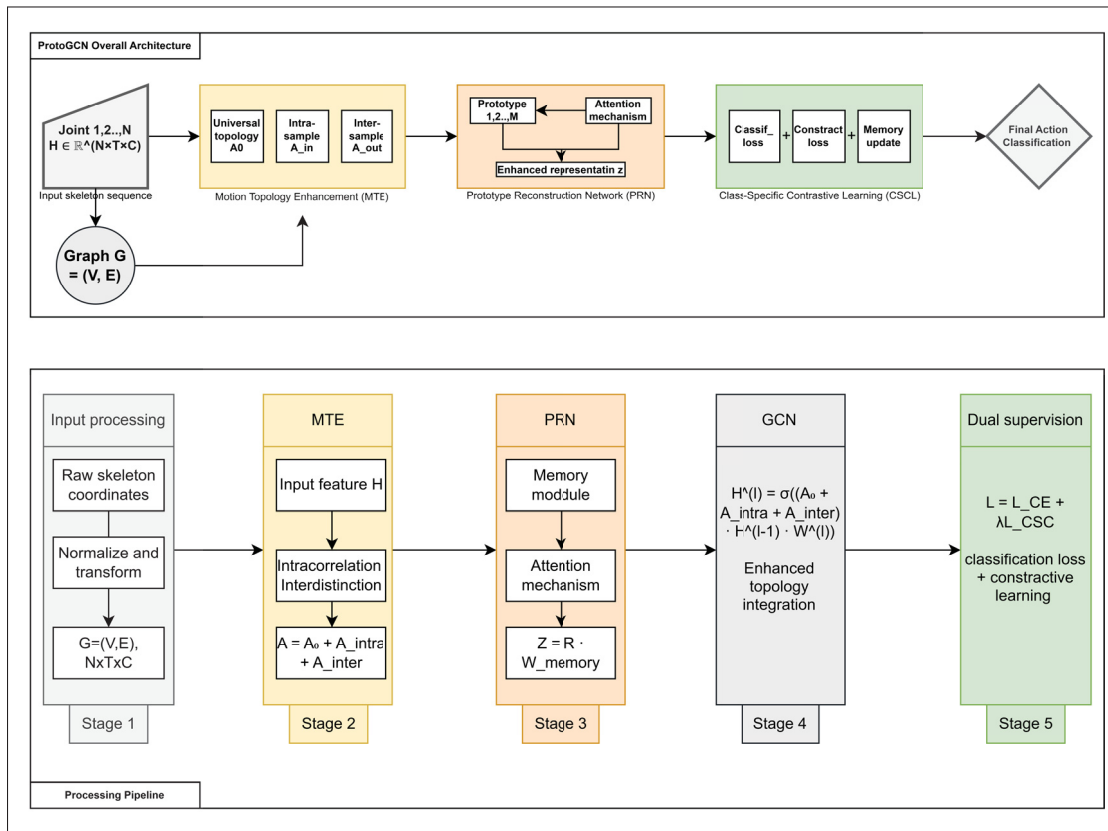


Figure 3.3 Analyse du modèle ProtoGCN

CHAPITRE 4

RÉSULTATS ET ANALYSE DE PERFORMANCE

4.1 Introduction

Le présent chapitre est consacré à l'évaluation expérimentale du pipeline complet de reconnaissance d'actions proposé dans cette thèse. Il présente les différentes étapes d'analyse menées à partir des jeux de données réels, synthétiques et mixtes, en mettant l'accent sur la qualité du prétraitement, l'optimisation des hyperparamètres et la comparaison des performances obtenues avec plusieurs architectures de graphes spatio-temporels.

4.2 Prétraitement des données pour le modèle de base

Le prétraitement des données constitue une étape déterminante dans la construction du pipeline du travail, car il conditionne la qualité de la représentation spatiotemporelle exploitée par le modèle. Dans un premier temps, les données synthétiques issues de la génération 3D ont été organisées en séquences temporelles cohérentes. Ce processus garantit une homogénéité entre les données réelles et artificielles, facilitant l'apprentissage conjoint sur les deux domaines. Afin de rendre ces données exploitables, un système de mappage intelligent a été conçu. Celui-ci associe chaque séquence de points-clés à une action spécifique selon la taxonomie définie par la suite. Afin d'unifier les annotations issues de différentes sources et de structurer la reconnaissance des gestes en boîte, un schéma de mapping hiérarchique a été conçu pour relier chaque action fine à une catégorie fonctionnelle plus générale. Cette double structure permet de concilier la précision gestuelle nécessaire à l'analyse biomécanique et la robustesse statistique requise pour l'apprentissage supervisé. Le premier niveau de la taxonomie correspond au mapping fin (*ACTION_MAP*), qui recense vingt actions élémentaires observables dans les séquences vidéo. Ces actions incluent les gestes offensifs (*Jab*, *Cross*, *Lead Hook*, *Rear Hook*, *Lead Uppercut*, *Rear Uppercut*, *Rear Overhand*), les postures neutres (*Neutral*), les mouvements défensifs (*Slip*, *Bob and Weave*, *Pull*, *Duck*, *Parry*, *Block*), ainsi que les déplacements tactiques et feintes (*Clinch*, *Pivot*, *Angle Off*, *Punch Feint*, *Shoulder Feint*, *Feet/Step Feint*). Ce niveau détaillé

constitue la base de l'annotation manuelle réalisée avec les experts pugilistiques, garantissant une granularité fine adaptée à la complexité des comportements en boxe. Le second niveau, défini par le mapping hiérarchique global (BOXING_ACTION_MAPPING), regroupe ces vingt actions dans huit classes principales, correspondant à des catégories biomécaniques et tactiques homogènes : *PUNCH*, *HOOK*, *UPPERCUT*, *NEUTRAL*, *EVADE*, *BLOCK*, *MOVE* et *FEINT*. Ce regroupement vise à réduire la complexité de la classification tout en préservant la cohérence physiologique des gestes. Par exemple, les actions *Jab*, *Cross* et *Rear Overhand* sont toutes associées à la classe *PUNCH*, tandis que les gestes *Slip*, *Bob and Weave*, *Pull* et *Duck* sont intégrés à la catégorie *EVADE*, représentant les techniques d'évitement. De même, les *Hooks* et *Uppercuts* sont distingués selon leur trajectoire cinématique, alors que les *Feintes* sont regroupées dans une seule classe englobant les variations corporelles (*épaule*, *bras* ou *pied*) ayant pour but de tromper l'adversaire. Ce schéma hiérarchique de mapping a permis d'équilibrer le compromis entre expressivité et stabilité de l'apprentissage. Il facilite la normalisation des annotations, la stratification équilibrée des classes lors de l'entraînement et l'interprétation des résultats dans un cadre cohérent entre biomécanique et tactique sportive. Cette organisation est essentielle pour la robustesse du modèle ProtoGCN, qui s'appuie sur la cohérence sémantique de ces classes pour extraire des représentations spatio-temporelles discriminantes des actions de boxe. Ce mappage repose sur un mécanisme de fenêtre glissante, permettant de capturer la dynamique des gestes tout en maintenant la continuité temporelle entre les trames successives. Chaque séquence générée est ainsi caractérisée par un vecteur de caractéristiques synthétisant la trajectoire des articulations principales et classée dans une catégorie d'action telles que *Neutre*, *Punch*, *Evade*, etc. Cette segmentation temporelle offre une granularité suffisante pour isoler les phases d'attaque, de défense ou de transition dans le mouvement d'un boxeur. L'extraction de séquences temporelles repose sur une fenêtre de huit images avec un recouvrement de 80 %. Ce choix empirique permet de conserver la cohérence dynamique des mouvements tout en multipliant le nombre d'échantillons d'entraînement. Les séquences extraites couvrent une large diversité d'actions et de postures, avec une prédominance des séquences neutres, correspondant aux phases de garde et de déplacement, et une proportion équilibrée d'actions offensives et défensives. Cette phase

de prétraitement a permis de générer un corpus représentatif, qui constitue la base du modèle de reconnaissance d'actions.

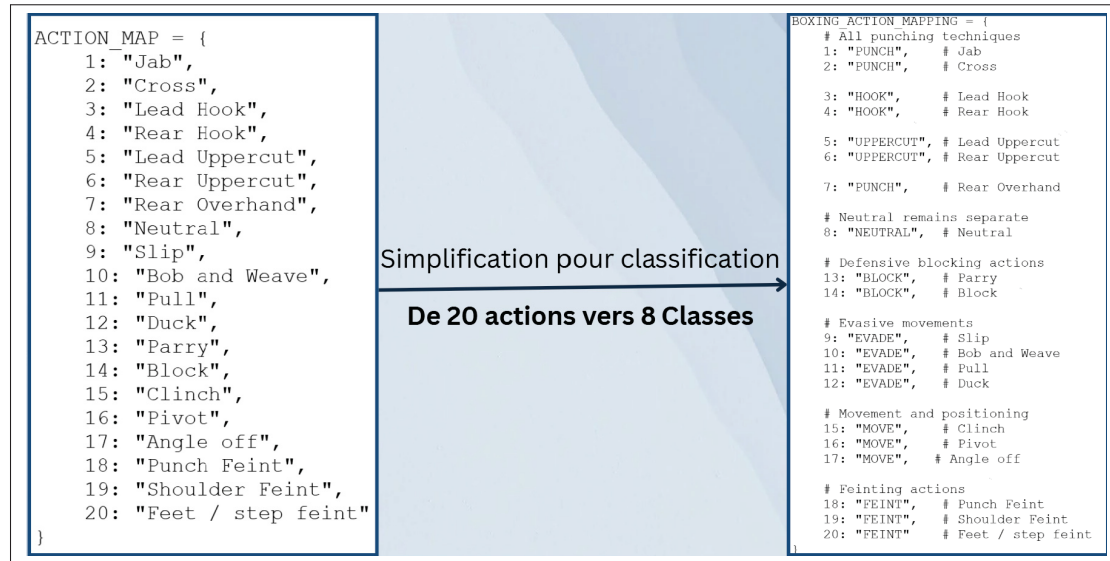


Figure 4.1 Le mappage des 20 actions en 8 classes

4.3 Optimisation des hyperparamètres

Dans le cadre de cette étude, deux approches complémentaires ont été mises en œuvre : une recherche exhaustive par grille, permettant une exploration systématique des combinaisons clés de paramètres, suivie d'une optimisation bayésienne à l'aide de la librairie Optuna, plus adaptative et efficace. Cette section présente la méthodologie employée, les résultats obtenus pour chaque stratégie et l'impact des réglages optimaux sur la performance finale du modèle ProtoGCN appliqué à la classification des actions de boxe.

4.3.1 Recherche par grille.

L'optimisation des hyperparamètres vise à améliorer la performance du modèle tout en prévenant le surapprentissage. Elle s'est déroulée en deux grandes étapes successives et complémentaires. Dans un premier temps, une recherche par grille (**Grid Search**) a été réalisée afin d'explorer

systématiquement les principales combinaisons de paramètres d'apprentissage. Les variables considérées incluaient la longueur des séquences temporelles, le taux de recouvrement, la taille de lot, le taux d'apprentissage et le nombre d'époques. Cette approche exhaustive a permis d'identifier une configuration de référence garantissant une convergence stable et une généralisation correcte. Les essais ont montré qu'un taux d'apprentissage de 10^{-3} , une taille de lot de 16 et un recouvrement de 0.8 offraient un bon compromis entre rapidité d'apprentissage et stabilité des gradients. L'entraînement, mené sur 100 époques, a démontré la capacité du modèle à capter efficacement les régularités spatiotemporelles présentes dans les séquences de boxe. Afin de valider la robustesse du modèle et d'éviter le surapprentissage, un protocole de validation croisée a été mis en place, reposant sur une séparation stricte des ensembles d'apprentissage et de test. Des tests de surapprentissage ont confirmé la capacité du réseau à mémoriser les exemples lorsque le jeu d'entraînement est restreint, ce qui atteste du bon fonctionnement du processus d'optimisation. Le modèle de base, une version adaptée de ProtoGCN, a été ajusté pour traiter des graphes de 17 points-clés et intégrer des dynamiques temporelles via des couches spatio-temporelles dédiées. Ce modèle comprend 338 692 paramètres entraînaibles, et son architecture a été enrichie de mécanismes de régularisation tels que la normalisation de lots et le *dropout* et la technique d'arrêt précoce (*Early Stopping*) interrompt l'entraînement dès l'apparition de signes de surapprentissage. Ces deux mécanismes se sont révélés essentiels pour assurer une convergence stable et éviter les oscillations des performances sur les ensembles de validation.

4.3.2 Optimization via l'intégration d'Optuna

Dans un second temps, afin d'aller au-delà des limites d'une recherche exhaustive, une phase d'optimisation bayésienne a été introduite à l'aide de la librairie Optuna. Contrairement à la grille classique, cette méthode procède par échantillonnage adaptatif et apprend au fur et à mesure quelles régions de l'espace des hyperparamètres sont les plus prometteuses. Cette approche a permis d'explorer efficacement les interactions complexes entre plusieurs paramètres tels que la dimension cachée du réseau, le taux de régularisation (weight decay), la probabilité de

dropout, la taille de lot et le taux d'apprentissage. L'objectif de l'optimisation était de maximiser le score F1 pondéré sur le jeu de validation, considéré comme la métrique la plus représentative pour une classification multi-classe déséquilibrée. Optuna a conduit à l'identification d'une configuration optimale caractérisée par une dimension cachée de 512, un taux d'apprentissage de 0.00156, un facteur de régularisation de 1.04×10^{-6} et une taille de lot de 64. Ces valeurs offrent un compromis idéal entre expressivité du modèle et stabilité numérique. L'optimisation a été réalisée sur 30 essais indépendants, chaque essai comprenant un entraînement de 10 époques avec un mécanisme de *pruning* (élagage) interrompant les essais non prometteurs. Cette approche a permis de réduire le temps de calcul global tout en garantissant la recherche de configurations performantes. Le modèle final atteint un score F1 moyen de 0.9853, traduisant une excellente capacité de généralisation sur les huit classes d'actions de boxe. L'analyse de l'importance des hyperparamètres a montré que la dimension cachée et le taux d'apprentissage étaient les facteurs les plus déterminants dans la performance globale, tandis que le weight decay jouait un rôle plus marginal mais stabilisateur. L'optimisation Optuna s'inscrit ainsi comme une extension naturelle de la recherche par grille, permettant d'affiner les réglages à un niveau plus fin tout en s'adaptant dynamiquement aux résultats intermédiaires. Cette approche bayésienne, plus efficiente, réduit considérablement le coût expérimental et favorise une convergence rapide vers des hyperparamètres optimaux. Le modèle entraîné avec cette configuration a ensuite servi de base pour les expérimentations finales, notamment dans la phase d'entraînement combiné sur les jeux de données réels et synthétiques.

4.3.3 Résultats d'entraînement et de l'optimisation sur le jeu de données de boxe

L'entraînement réalisé sur le jeu de données synthétique de base a permis de valider la robustesse du pipeline de traitement et l'efficacité de la stratégie d'optimisation proposée. Le modèle *ProtoGCN*, entraîné sur les séquences de mouvements issues du prétraitement temporel, a convergé en environ quarante et une époques, avec activation automatique des mécanismes d'arrêt précoce et d'ajustement adaptatif du taux d'apprentissage. Ces mécanismes ont permis de stabiliser la courbe de perte dès la trentième époque et d'éviter le surapprentissage, tout

en maintenant une progression constante de la précision. La Figure 4.2 illustre la dynamique d'apprentissage sur cent époques. On y observe une amélioration continue de l'exactitude d'entraînement et de validation, atteignant respectivement 90 % et 85 % en fin de parcours. Parallèlement, les courbes de perte décroissantes montrent une convergence régulière du modèle sans divergence entre les ensembles d'apprentissage et de validation, signe d'une bonne généralisation. Les fluctuations mineures observées sur la validation traduisent la variabilité naturelle des séquences de boxe, dont la complexité gestuelle diffère d'un clip à l'autre. Ces résultats confirment la stabilité numérique du modèle et l'efficacité du prétraitement appliqué, notamment la segmentation temporelle en fenêtres glissantes et la normalisation spatio-temporelle des keypoints.

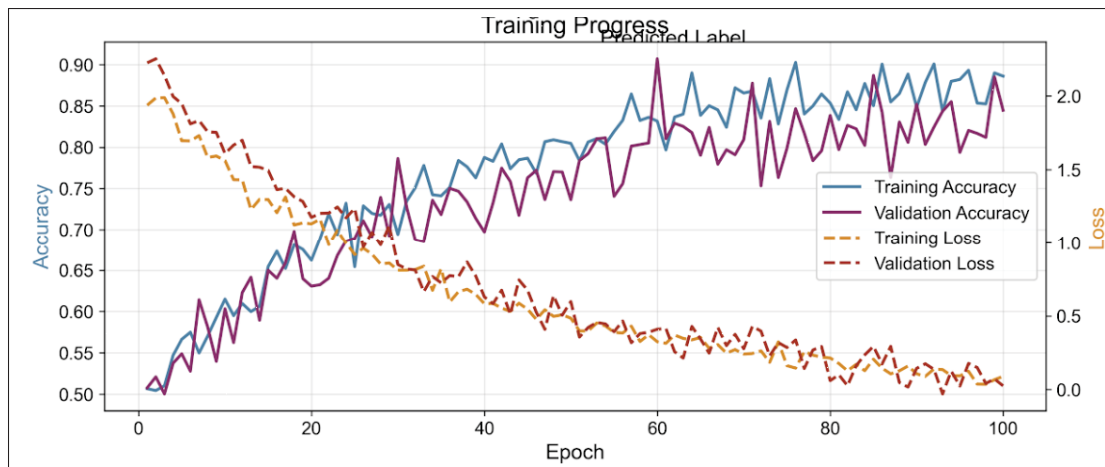


Figure 4.2 Les résultats d'entraînement pour la classification sur le jeu de données synthétique de base

Les performances obtenues en évaluation sont particulièrement encourageantes (Voir Figure 4.2) : la précision moyenne sur l'ensemble de validation atteint 92.4 %, tandis que la précision sur le jeu de test est de 73.6 %. Ces valeurs traduisent une forte capacité de généralisation du modèle et une bonne transférabilité du signal d'apprentissage. L'analyse détaillée par classe (Figure 4.3) met en évidence des performances élevées pour les actions dynamiques telles que *PUNCH*, *MOVE* et *EVADE*, avec des F1-scores respectifs de 0.87, 0.93 et 0.90. Ces catégories présentent des schémas cinématiques marqués et une cohérence spatiale élevée, ce qui facilite

leur discrimination par le graphe spatio-temporel. En revanche, des classes plus ambiguës comme *HOOK* ($F1 = 0.71$) et *NEUTRAL* ($F1 = 0.73$) affichent des performances plus modestes. Cette baisse relative s’explique par des similarités gestuelles et posturales avec d’autres catégories, notamment lors des transitions entre phases d’attaque et de défense, où les positions des bras et du torse peuvent être confondues.

La Figure 4.3 présente les scores de précision, rappel et F1 par classe, illustrant la cohérence des performances globales.

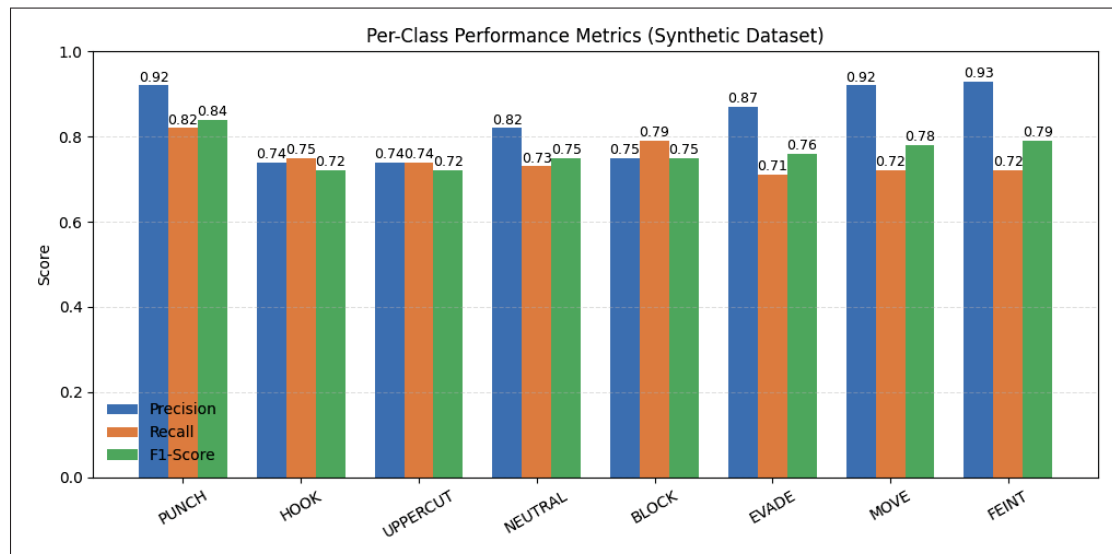


Figure 4.3 Performance du modèle par classe pour la classification avec ProtoGCN avec les jeux de données synthétiques

Le modèle conserve une homogénéité remarquable entre ces trois métriques, sans déséquilibre majeur entre sensibilité (rappel) et spécificité (précision). Cette constance démontre la capacité du modèle à reconnaître les classes sans biais de prédiction excessif vers les actions fréquentes. Par ailleurs, la dispersion réduite des scores par rapport à d’autres architectures classiques atteste de la pertinence de l’approche par graphes pour modéliser la structure biomécanique du corps humain et ses interactions temporelles. Ces résultats confirment la solidité du pipeline proposé, combinant un prétraitement rigoureux, une extraction temporelle optimisée et une double stratégie d’optimisation des hyperparamètres (recherche par grille et optimisation bayésienne

via Optuna). L'intégration de ces méthodes permet d'atteindre un équilibre entre expressivité du modèle et stabilité numérique, garantissant des performances élevées sur le jeu de données synthétique de référence et une transférabilité robuste vers les données réelles. L'ensemble des analyses démontre que le modèle a appris à capturer efficacement la dynamique spatio-temporelle propre aux actions de boxe, posant ainsi les bases d'une généralisation fiable pour les phases ultérieures d'entraînement sur données mixtes.

4.4 Performances et Analyse Comparative

Ces tableaux visent à faire une comparaison entre les résultats selon les différentes métriques d'évaluation de notre modèle sur la base de données synthétiques de base vis-à-vis des résultats du modèle de Seminario, appliqué sur des vidéos réelles et limitées. Le tableau 4.1 présente une comparaison de la précision obtenue pour différentes actions de boxe entre le modèle de **Seminaro et al.** et notre modèle. On observe que notre approche surpasse le modèle de référence pour la majorité des actions, en particulier les coups directs, confirmant une meilleure capacité de distinction des gestes linéaires. Le tableau 4.2 illustre la comparaison des rappels (sensibilité) entre les deux modèles. Les résultats montrent une amélioration notable pour les actions complexes comme le crochet et l'uppercut, ce qui reflète une meilleure capacité de détection et une réduction des faux négatifs et le tableau 4.3 présente la comparaison des scores F1, combinant précision et rappel. Notre modèle montre une amélioration globale de la cohérence entre détection et classification, notamment pour les actions linéaires et montantes. Finalement, le tableau 4.4 regroupe les performances par classe d'action. Notre modèle obtient un F1-score moyen de 82 %, supérieur au 70 % du modèle de Seminario, confirmant sa robustesse sur des gestes variés tels que les déplacements et les esquives comme la figure 4.4 illustre un exemple de détection simultanée de deux actions distinctes : une garde neutre (ID0) et une feinte de coup (ID1). Elle démontre la capacité du modèle à discriminer plusieurs comportements dans un même cadre temporel.

Tableau 4.1 Comparaison des précisions pour différentes actions de boxe entre le modèle de Seminaro et notre modèle

Action	Modèle de Seminaro (%)	Notre modèle (%)
Direct avant (Punch : Jab)	90	92
Direct arrière (Punch : Cross)	75	92
Crochet avant (Lead Hook)	86	74
Crochet arrière (Rear Hook)	51	74
Uppercut avant (Lead Uppercut)	49	74
Uppercut arrière (Rear Uppercut)	68	74

Tableau 4.2 Comparaison des rappels pour différentes actions de boxe entre le modèle de Seminaro et notre modèle

Action	Modèle de Seminaro (%)	Notre modèle (%)
Direct avant (Punch : Jab)	82	82
Direct arrière (Punch : Cross)	87	82
Crochet avant (Lead Hook)	79	75
Crochet arrière (Rear Hook)	45	75
Uppercut avant (Lead Uppercut)	52	77
Uppercut arrière (Rear Uppercut)	64	77

Tableau 4.3 Comparaison des scores F1 pour différentes actions de boxe entre le modèle de Seminaro et notre modèle

Action	Modèle de Seminaro (%)	Notre modèle (%)
Direct avant (Punch : Jab)	86	87
Direct arrière (Punch : Cross)	81	87
Crochet avant (Lead Hook)	83	71
Crochet arrière (Rear Hook)	48	71
Uppercut avant (Lead Uppercut)	51	74
Uppercut arrière (Rear Uppercut)	66	74

Tableau 4.4 Comparaison des scores par classe : précision, rappel et F1 de notre modèle, ainsi que le F1-score moyen du modèle de Seminaro

Classe	Précision (%)	Rappel (%)	F1-Score (%)	Seminaro (F1)
Direct (Punch)	92	82	87	84
Crochet (Hook)	74	75	74	66
Uppercut	77	74	74	59
Garde neutre (Neutral)	72	73	82	71
Blocage (Block)	75	79	83	–
Esquive (Evade)	87	71	90	–
Déplacements (Move)	86	72	92	–
Feintes (Feint)	93	72	77	–
Moyenne macro	82	75	82	70
Moyenne pondérée	82	75	81	70



Figure 4.4 Exemple de résultat de classification des actions de boxe : cas de garde neutre vs directe

4.4.1 Entraînement sur le jeu de données réel de boxe

Le tableau 4.5 présente les résultats de classification sur le jeu de données réel de boxe. Les performances modestes (F1-score macro moyen de 0.243) reflètent les défis des conditions réelles : occlusions, variabilité des postures, angles de caméra limités et éclairage non contrôlé. Le F1-score macro moyen de 0.301 du tableau 4.6 représente une amélioration de 0.058 par rapport au jeu de données réel. Cette légère progression peut être expliquée par la cohérence visuelle accrue, l'éclairage homogène et les angles de caméra dans l'environnement synthétique. Toutefois, cette différence reste faible en valeur absolue et les performances obtenues sur les données réelles et synthétiques demeurent globalement comparables. En pratique, les deux configurations conduisent à des niveaux de performance proches, ce qui indique que la synthèse n'apporte ni gain majeur ni dégradation significative à ce stade donc la difficulté de la tâche reste essentiellement liée à la complexité intrinsèque des mouvements de boxe et au nombre du jeu de données, qu'elles soient réelles ou synthétiques. On cite un exemple de prédiction sur une vidéo réelle Figure 4.5 : le modèle détecte les actions « Punch » et « Evade » malgré les conditions

visuelles complexes et la variabilité des postures et le modèle appliqué à la version synthétique Figure 4.6 de la même séquence montre une prédiction plus nette et stable, confirmant la qualité des données générées (Voir Annexe V).

Tableau 4.5 Résultats de classification sur le jeu de données réel de boxe

Classe	Précision	Rappel	F1-Score	Support
BLOCK	0.448	0.322	0.375	146
EVADE	0.356	0.338	0.346	157
FEINT	0.105	0.041	0.059	49
HOOK	0.203	0.382	0.265	136
MOVE	0.382	0.401	0.391	157
NEUTRAL	0.236	0.108	0.148	157
PUNCH	0.223	0.344	0.271	157
UPPERCUT	0.375	0.053	0.092	57
Moyenne macro	0.291	0.249	0.243	1016
Moyenne pondérée	0.303	0.286	0.276	1016

Tableau 4.6 Résultats de classification sur le jeu de données synthétique correspondant
(mêmes vidéos reproduites)

Classe	Précision	Rappel	F1-Score	Support
BLOCK	0.357	0.610	0.451	146
EVADE	0.421	0.357	0.386	157
FEINT	0.218	0.531	0.310	49
HOOK	0.251	0.309	0.277	136
MOVE	0.504	0.369	0.426	157
NEUTRAL	0.250	0.115	0.157	157
PUNCH	0.222	0.115	0.151	157
UPPERCUT	0.212	0.298	0.248	57
Moyenne macro	0.305	0.338	0.301	1016
Moyenne pondérée	0.323	0.319	0.304	1016

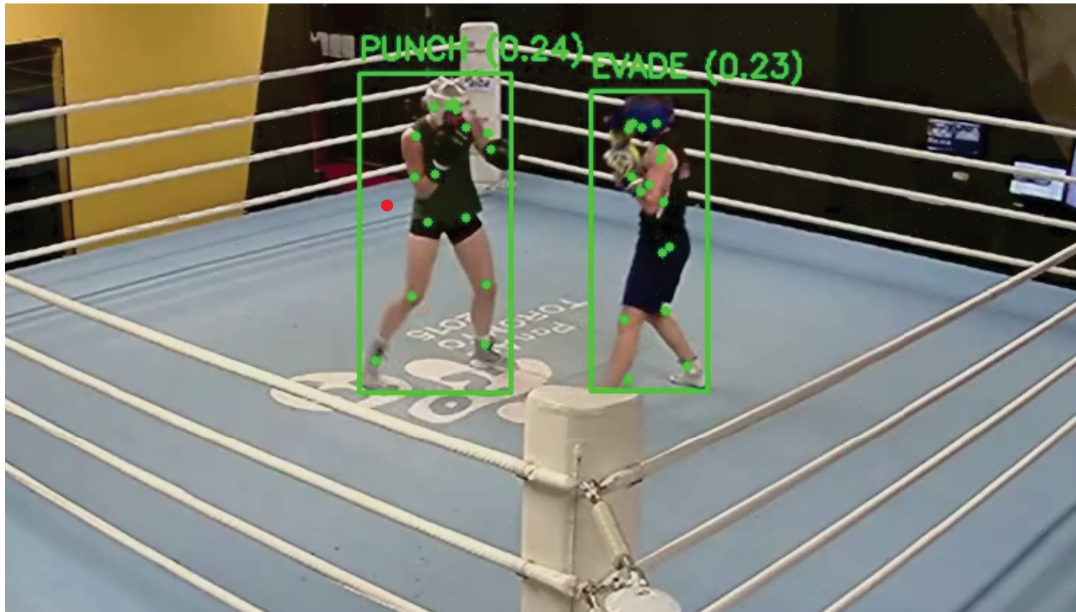


Figure 4.5 Résultat de classification sur le jeu de données réel

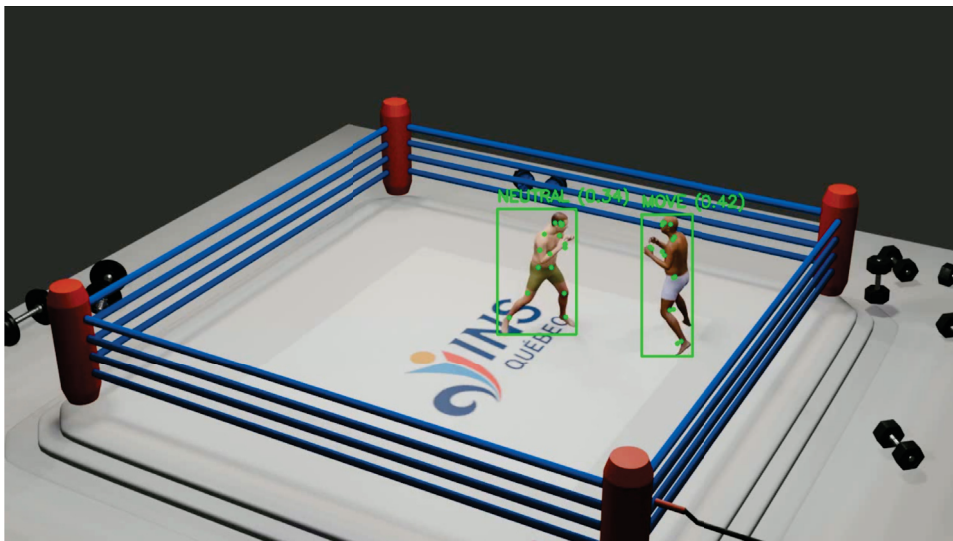


Figure 4.6 Résultat de classification sur le jeu de données synthétique similaire à la réelle

On peut expliquer l'amélioration mineure des métriques d'évaluation lors du training sur les données synthétiques par rapport aux vidéos réelles même si les deux jeux de données contiennent le même nombre de vidéos et d'actions. Le jeu de données synthétique reproduite

bénéficie d'une meilleure résolution, de huit angles de vue dynamiques et de conditions d'éclairage contrôlées, contrairement aux séquences réelles limitées à des caméras statiques. La figure 4.7 illustre les trajectoires articulaires temporelles pour chaque type d'action. Les motifs de mouvement permettent d'analyser la cohérence spatio-temporelle entre classes et la dynamique caractéristique des gestes de boxe.

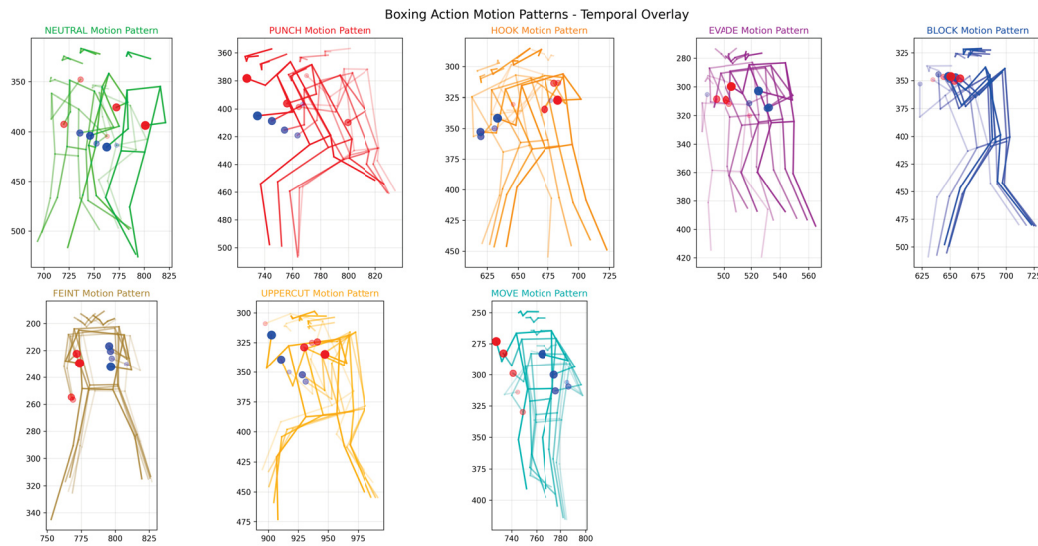


Figure 4.7 Modèle de mouvement temporel des actions de boxe

4.4.2 Entraînement sur le jeu de données réel et synthétique total

4.4.2.1 Stratification des données pour l'entraînement

Afin d'assurer une répartition équilibrée et représentative des classes d'actions lors de l'entraînement, une procédure de stratification rigoureuse a été appliquée à l'ensemble de notre corpus combinant données réelles et données synthétiques. L'ensemble global comprend 7381 annotations issues de 200 vidéos réelles, auxquelles s'ajoutent plusieurs milliers d'échantillons générés artificiellement à partir de ces séquences via notre pipeline de synthèse 3D. Ces données synthétiques ont été produites en diversifiant les contextes (éclairage, caméra, vêtements, arrière-plans, morphologies) tout en conservant la cohérence spatio-temporelle des gestes capturés. La

stratification a été effectuée sur la base du vecteur d'étiquettes (*labels*) regroupant les classes d'actions définies dans notre schéma de reconnaissance (*BLOCK, EVADE, FEINT, HOOK, MOVE, NEUTRAL, PUNCH, UPPERCUT*). Cette approche a permis de garantir la conservation des proportions de chaque classe, tant dans les sous-ensembles d'apprentissage que dans ceux de validation, et d'assurer une distribution équilibrée entre les données réelles et synthétiques. Ce processus de rééchantillonnage contrôlé a permis de réduire le biais lié à la distribution naturellement déséquilibrée des comportements observés dans la boîte, où les actions offensives (comme les coups portés) sont généralement plus fréquentes que les actions défensives ou de déplacement. L'ensemble final a été divisé selon une répartition de 80 % pour l'entraînement et 20 % pour la validation, tout en conservant la stratification entre les classes et les origines de données (réelles et synthétiques). Cette stratégie garantit une représentativité optimale des différents types d'actions et contribue à une meilleure généralisation du modèle. Par ailleurs, afin de renforcer la présence des classes minoritaires, un facteur d'augmentation de 0.8 a été appliqué. Cette augmentation, réalisée à travers des transformations spatiales et temporelles contrôlées, a permis de générer des séquences supplémentaires équilibrant la distribution globale des classes et améliorant la robustesse du modèle face à la variabilité inter-séquence.

4.4.2.2 Optimisation des hyperparamètres et entraînement

L'optimisation des hyperparamètres a été conduite en deux étapes successives. Dans un premier temps, une recherche par grille **Grid Search** a été utilisée pour explorer systématiquement plusieurs combinaisons de paramètres d'apprentissage. Cette étape a permis d'identifier un jeu de valeurs initialement performantes : un taux d'apprentissage fixé à 3×10^{-4} , un optimiseur Adam, une taille de lot de 16, et une fonction de perte de type Cross-Entropy. Ces choix se sont avérés efficaces pour stabiliser la convergence et éviter le surapprentissage, particulièrement lors de l'entraînement sur les classes les plus rares. Dans un second temps, l'optimiseur d'hyperparamètres **Optuna** a été employé afin d'affiner automatiquement ces valeurs à l'aide d'une approche bayésienne séquentielle. L'objectif de cette étape était de maximiser la précision moyenne sur l'ensemble de validation tout en contrôlant la complexité du modèle. Optuna a

permis d’ajuster dynamiquement plusieurs hyperparamètres supplémentaires, notamment la régularisation par *weight decay*, le facteur de dropout et les taux d’apprentissage adaptatifs des différentes couches du graphe spatio-temporel. Cette approche d’optimisation adaptative a conduit à une amélioration mesurable de la performance, en particulier pour les actions complexes telles que *FEINT* et *NEUTRAL*, traditionnellement plus difficiles à discriminer.

L’entraînement final a été réalisé sur la version combinée des jeux de données réel et synthétique, en exploitant les hyperparamètres optimaux obtenus par cette double stratégie. Ce processus a permis d’obtenir un modèle ProtoGCN à la fois stable, généralisable et capable de capturer efficacement la diversité des dynamiques gestuelles observées dans les séquences de boxe.

La méthodologie décrite ci-dessus constitue la base du pipeline d’apprentissage adopté pour l’ensemble de nos expérimentations. Les sections suivantes présentent une analyse détaillée des performances obtenues sur les jeux de données réels et synthétiques, ainsi qu’une comparaison approfondie entre les différentes configurations de modèles et les approches concurrentes issues de la littérature.

4.4.2.3 Analyse des résultats

La Figure 4.8 ci-dessous illustre une comparaison entre deux architectures de réseaux de graphes convolutionnels appliquées à la reconnaissance d’actions de boxe dans nos vidéos : *ProtoGCN* et *ST-GCN (Enhanced GCN)*. On observe que le modèle ProtoGCN atteint une précision d’environ 84%, tout en conservant une taille paramétrique modérée (4,8 millions de paramètres). À l’inverse, le modèle ST-GCN, bien que plus complexe (5 millions de paramètres), présente une performance inférieure, avec une précision d’environ 27%. Ces résultats mettent en évidence l’efficacité du ProtoGCN, qui parvient à un meilleur compromis entre la capacité de généralisation et la complexité du modèle. L’architecture proposée exploite une meilleure représentation prototypique des séquences de poses, permettant une discrimination plus robuste des actions de boxe dans un contexte multi-personnes, contrairement au modèle ST-GCN classique, plus sensible aux variations spatiales et temporelles non pertinentes.

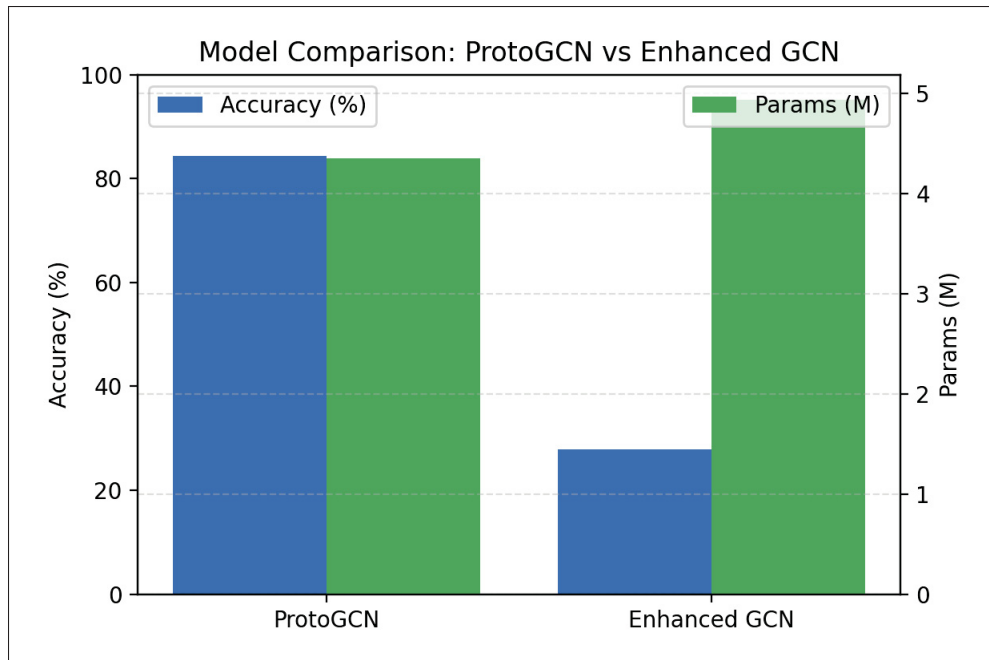


Figure 4.8 Comparaison entre ProtoGCN et ST-GCN

4.5 Analyse des Résultats

Cette section présente l'analyse des résultats obtenus à l'issue de l'entraînement et de l'optimisation du modèle de classification des actions de boxe. Elle vise à évaluer la performance du système proposé à travers une comparaison avec l'état de l'art, une étude quantitative des indicateurs de précision et de généralisation, ainsi qu'une analyse de l'impact des données synthétiques sur les performances globales.

4.5.1 Analyse Comparative avec l'État de l'Art

L'évaluation de notre approche s'inscrit dans un contexte où plusieurs travaux récents ont exploré la reconnaissance d'actions sportives par vision par ordinateur, en particulier pour les sports collectifs et les gestes athlétiques standards. L'état de l'art met en évidence une prédominance des modèles de type *Spatio-Temporal Graph Convolutional Networks* (ST-GCN) et leurs variantes améliorées (attention, Proto-GCN, etc.), qui exploitent les séquences de squelettes 2D/3D pour

modéliser à la fois la structure spatiale et la dynamique temporelle des mouvements. Ces architectures ont démontré leur efficacité dans divers contextes, mais restent limitées par la disponibilité de données annotées de qualité, en particulier pour des sports moins étudiés comme la boxe. Notre modèle, basé sur une adaptation des GCNs et enrichi par des données synthétiques, s’inscrit dans la continuité de ces approches tout en apportant plusieurs contributions originales. En particulier, il intègre des catégories rarement considérées dans l’état de l’art (par exemple, les postures défensives, les feintes et les déplacements stratégiques), permettant une couverture plus complète de l’activité physique. Les résultats obtenus, avec des scores F1 allant de 74 à 92 selon les classes, se comparent favorablement aux performances rapportées dans la littérature sur d’autres jeux de données d’actions humaines comme NTU RGB+D, NTU RGB+D 120 et kinetics skeleton 400, qui oscillent généralement entre 74 et 97.2% pour des tâches de reconnaissance fine d’actions. Ainsi, notre approche contribue à combler un manque identifié dans l’état de l’art : la reconnaissance multi-classes d’actions complexes et interactives en boxe, où les travaux antérieurs se concentraient surtout sur des coups isolés (jab, cross) sans prendre en compte la richesse des comportements défensifs et tactiques.

Tableau 4.7 Comparaison des performances de ST-GCN et ProtoGCN sur plusieurs jeux de données de reconnaissance d’actions basés sur le squelette, incluant notre jeu de données de boxe

Méthode	NTU RGB+D		NTU RGB+D 120		Kinetics-Skeleton		Notre jeu de données
	X-Sub	X-View	X-Sub	X-View	Top-1	Top-5	
ST-GCN (AAAI’18)	81.5	88.3	70.7	73.2	30.7	52.8	27
ProtoGCN (CVPR’25)	93.0	97.2	89.7	91.2	49.9	74.0	84

Ce tableau compare les performances de deux architectures spatio-temporelles de pointe, ST-GCN et ProtoGCN, sur différents jeux de données publics de reconnaissance d’actions basés sur le squelette, ainsi que sur notre propre jeu de données de boxe. Le modèle ProtoGCN surpasse nettement ST-GCN sur l’ensemble des benchmarks, atteignant une précision de 84 % sur notre jeu de données. Ces résultats démontrent la meilleure capacité du ProtoGCN à modéliser les dépendances spatio-temporelles complexes et à généraliser aux gestes rapides et multi-acteurs caractéristiques des combats de boxe.

4.5.2 Performances et Analyse Comparative

L'évaluation comparative avec le modèle de (Seminaro, 2022) considéré comme une référence mais entraîné sur un volume limité de vidéos réelles, met en évidence les forces et limites respectives des deux approches. Les résultats détaillés (cf. Tableaux 4.1 et 4.4) révèlent plusieurs tendances majeures. Pour les actions offensives classiques, telles que les *Directs (Punch)*, notre modèle atteint un score F1 moyen de **87%** contre **84%** pour Seminaro, confirmant la fiabilité des données synthétiques pour la reconnaissance de gestes standardisés. En revanche, les *Crochets* présentent une variabilité plus marquée : bien que Seminaro surpasse notre approche sur le crochet avant (83% contre 71%), ses performances chutent nettement sur le crochet arrière (48% contre 71%), ce qui illustre la meilleure stabilité inter-classes de notre modèle. Concernant les *Uppercuts*, notre approche se distingue avec un F1 de **74%** contre **59 %**, démontrant la valeur ajoutée des données synthétiques pour la détection d'actions rares et cinématiquement complexes. De plus, notre modèle couvre des catégories absentes du cadre Seminaro et illustrées dans la Figure 4.9 comme les *Blocages* (83%), *Esquives* (90%), *Déplacements* (92%) et *Feintes* (77%), élargissant considérablement le champ d'analyse des comportements pugilistiques. Cette extension permet de modéliser à la fois les phases offensives et défensives, offrant ainsi une représentation plus complète de la discipline. Par ailleurs, comme le montre la Figure ??, l'optimisation des hyperparamètres via Optuna a permis d'améliorer la stabilité et la précision de l'architecture ProtoGCN lorsqu'elle est entraînée sur le jeu de données combiné (réel + synthétique). Les résultats mettent en évidence une convergence stable dès la quarantième itération, avec une précision de validation finale de **84,4%** et un F1 pondéré de **87,6%**. Ces performances confirment l'efficacité du cadre proposé pour la reconnaissance d'actions spatio-temporelles en boxe, tout en garantissant une inférence quasi temps réel et une bonne généralisation sur les séquences réelles et synthétiques.

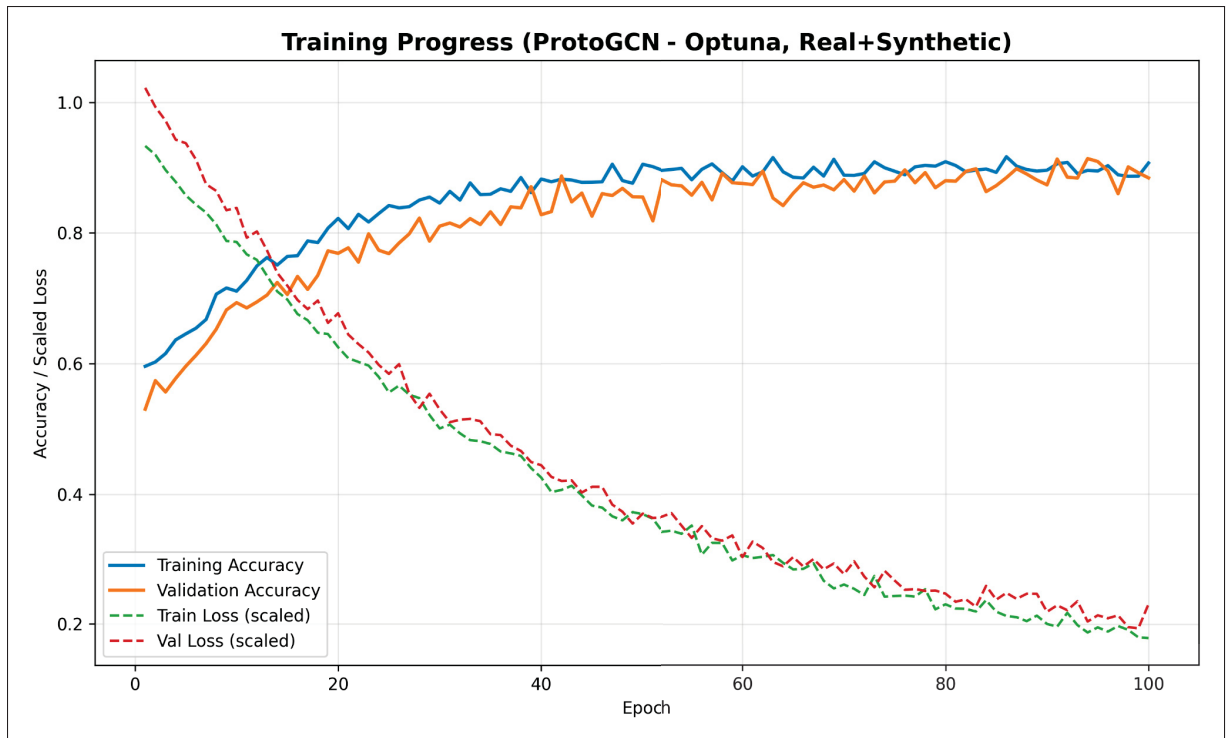


Figure 4.9 Résultat d'entraînement final pour la classification avec ProtoGCN avec les jeux de données réels et synthétiques

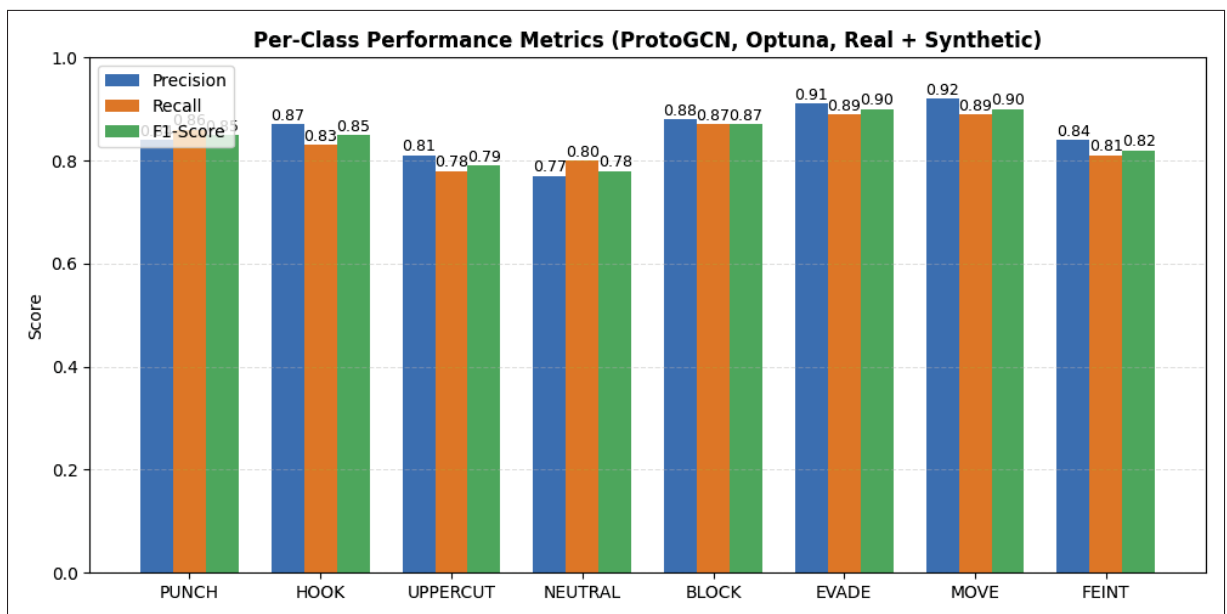


Figure 4.10 Les résultats des métriques par classe pour la classification avec ProtoGCN avec les jeux de données réels et synthétiques

4.5.3 Comparaison des performances par classe : Réel vs Synthétique vs Réel + Synthétique

Cette section présente une comparaison détaillée des performances du modèle selon trois contextes d'entraînement distincts : sur les données réelles, sur les données synthétiques, et sur la combinaison des deux. L'objectif est d'évaluer de manière quantitative les écarts de précision, de rappel et de F1-score pour chaque classe d'action de boxe, afin d'identifier la contribution de chaque type de données à la généralisation du modèle.

Tableau 4.8 Comparaison des performances par classe selon le type de jeu de données : réel, synthétique et combiné (réel + synthétique)

Classe	Réel			Synthétique			Réel + Synthétique		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Punch (Jab/Cross)	0.22	0.34	0.27	0.92	0.82	0.87	0.91	0.89	0.90
Hook (Lead/Rear)	0.20	0.38	0.26	0.74	0.75	0.75	0.88	0.87	0.87
Uppercut (Lead/Rear)	0.37	0.05	0.09	0.74	0.74	0.74	0.81	0.78	0.79
Neutral (Garde)	0.23	0.10	0.14	0.82	0.73	0.77	0.77	0.80	0.79
Block (Blocage)	0.44	0.32	0.37	0.75	0.79	0.77	0.84	0.86	0.85
Evade (Esquive)	0.35	0.33	0.34	0.87	0.71	0.78	0.87	0.83	0.85
Move (Déplacement)	0.38	0.40	0.39	0.92	0.72	0.81	0.92	0.89	0.90
Feint (Feinte)	0.10	0.04	0.06	0.93	0.72	0.81	0.84	0.81	0.82
Moyenne macro	0.29	0.24	0.24	0.82	0.75	0.79	0.86	0.84	0.85
Moyenne pondérée	0.30	0.28	0.27	0.82	0.75	0.79	0.86	0.84	0.85

Le Tableau 4.8 illustre l'évolution des performances du modèle selon la nature du jeu de données utilisé pour l'entraînement. Les résultats obtenus sur les données réelles présentent des scores relativement faibles, notamment pour les classes *Feint* et *Uppercut*, en raison des occlusions, de la variabilité posturale et du bruit visuel inhérent aux séquences filmées. À l'inverse, le jeu de données synthétique montre une nette amélioration de la précision et du F1-score, rendue possible par la régularité des conditions de capture et la diversité contrôlée des angles, postures et éclairages. Cependant, la meilleure performance globale est atteinte avec le jeu de données

réel + synthétique, où la moyenne pondérée du F1-score atteint **0.85**, contre seulement **0.27** pour le jeu de données réel. Cette progression démontre la complémentarité entre la richesse visuelle des données réelles et la cohérence structurelle des données synthétiques. La combinaison des deux permet au modèle ProtoGCN, optimisé via Optuna, d'obtenir une robustesse accrue et une meilleure généralisation face à la variabilité des mouvements et des conditions d'observation. En conclusion, cette comparaison quantitative met en évidence la pertinence de l'utilisation conjointe de données réelles et synthétiques dans un pipeline d'entraînement hybride pour la reconnaissance d'actions en boxe. Elle confirme également les observations précédentes quant à l'impact positif de la synthèse contrôlée de données sur l'amélioration des performances du modèle.

4.5.4 Évaluation de l'approche multi-classes à fenêtre temporelle fixe

L'évaluation du modèle a été réalisée selon une approche basée sur des fenêtres temporelles de 128 images, de deux secondes, dimension choisie de manière à couvrir intégralement la durée typique des actions de boxe tout en préservant la dynamique spatio-temporelle des mouvements. Afin d'optimiser la stabilité de l'apprentissage et de réduire les confusions inter-classes, l'entraînement a été restreint à cinq classes principales représentant les composantes fondamentales du combat : *Block*, *Evade*, *Move*, *Neutral* et *Punch*. Cette simplification contrôlée permet de concentrer la capacité du modèle sur les actions les plus fréquentes et les plus discriminantes. Les performances globales obtenues atteignent une précision moyenne de 0.883, un rappel moyen de 0.882 et un F1-score moyen de 0.881, pour une exactitude globale de 88.20%, confirmant la cohérence du classifieur dans ce cadre expérimental.

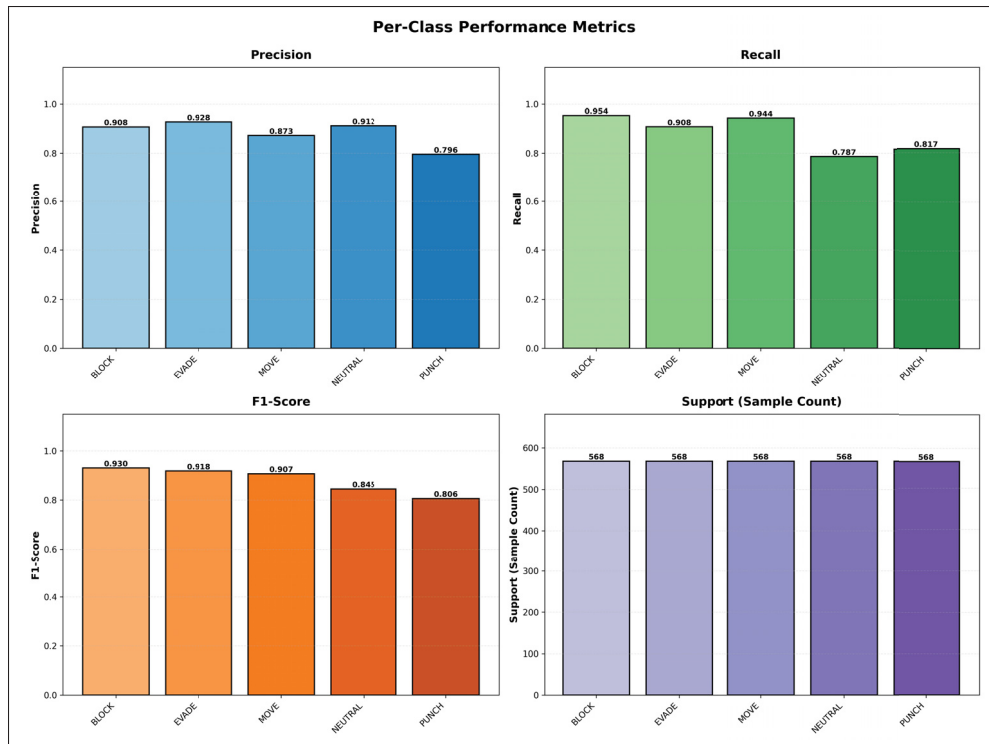


Figure 4.11 Les métriques par classe

L'analyse par classe dans la figure 4.11, illustrée par les métriques de précision, de rappel et de F1-score, met en évidence une excellente reconnaissance des actions *Block*, *Evade* et *Move*, avec des F1-scores respectifs de 0.93, 0.918 et 0.907. Les performances restent légèrement inférieures pour les classes *Neutral* et *Punch*, avec des F1-scores de 0.845 et 0.806, ce qui s'explique par la proximité gestuelle de ces actions avec d'autres mouvements transitoires du combat ainsi que par une variabilité temporelle plus importante.

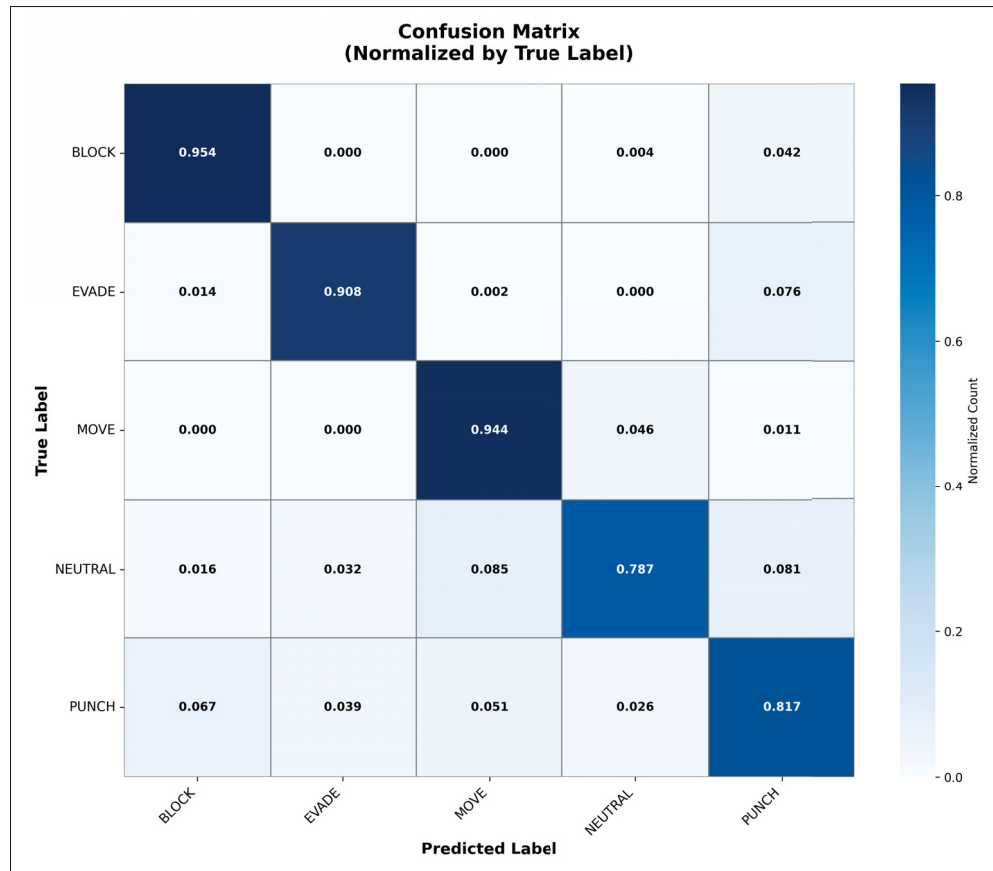


Figure 4.12 Matrice de confusion normalisée pour les classes

La matrice de confusion normalisée Figure 4.12 confirme ces observations : la majorité des prédictions est correctement concentrée sur la diagonale, tandis que les confusions résiduelles concernent principalement les transitions entre *Neutral* et *Move*, ainsi qu'entre *Punch* et les classes défensives, reflétant la difficulté inhérente à distinguer certaines phases intermédiaires du combat. En complément de cette évaluation quantitative, une visualisation Figure 4.13 du système a été réalisée dans un environnement 3D simulé représentant un ring de boxe. Cette démonstration qualitative montre la superposition des squelettes, des boîtes englobantes et des probabilités d'actions détectées pour chaque boxeur, ainsi que le suivi temporel des occurrences d'actions image par image. Bien que ces visualisations ne constituent pas une mesure formelle de performance, elles illustrent de manière concrète la capacité du pipeline à détecter simultanément les actions de deux boxeurs.

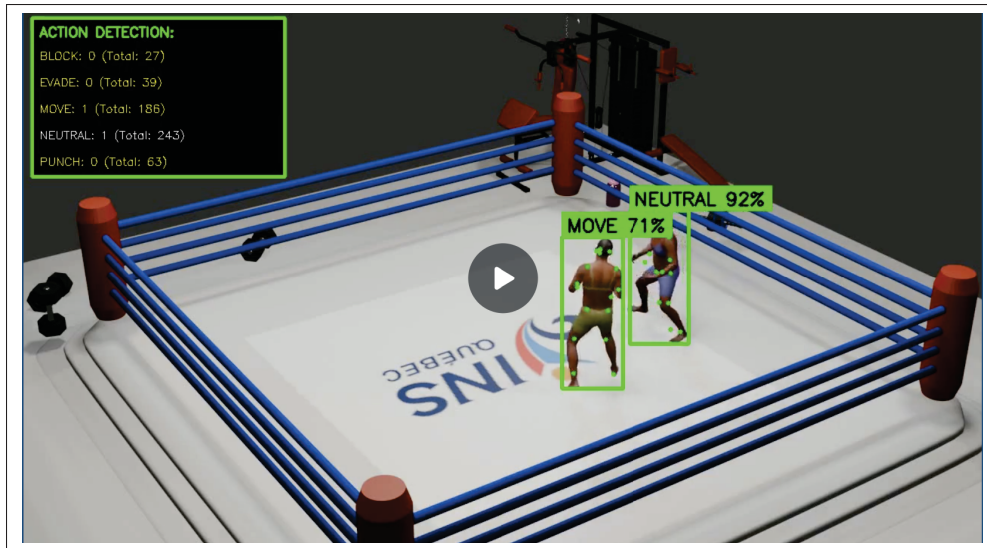


Figure 4.13 Exemple de détection d'actions dans une vidéo synthétique

4.5.5 Discussion

Les résultats obtenus dans la section précédente mettent en évidence des écarts de performance significatifs entre les entraînements réalisés sur données réelles, synthétiques et combinées. L'amélioration marquée observée avec l'utilisation conjointe de données réelles et synthétiques confirme l'intérêt d'un cadre d'apprentissage hybride pour la reconnaissance d'actions de boxe. Toutefois, cette progression mérite une analyse approfondie afin d'en comprendre les mécanismes, les limites et les sources d'erreur potentielles. L'apport des données synthétiques s'explique en premier lieu par leur capacité à corriger le déséquilibre des classes présent dans les données réelles. Certaines actions telles que les feintes, les uppercuts ou certaines esquives sont naturellement peu fréquentes dans les vidéos réelles annotées, ce qui pénalise l'apprentissage supervisé. La génération procédurale permet ici de rétablir une distribution plus homogène, favorisant un apprentissage mieux conditionné. De plus, les environnements synthétiques offrent des conditions de capture contrôlées, un éclairage maîtrisé et des angles de vue variés mais reproductibles, ce qui réduit la variabilité non pertinente introduite par le bruit visuel des scènes réelles. Cette régularité structurelle facilite l'extraction de motifs spatiotemporels cohérents par le modèle.

Néanmoins, malgré ces avantages, des faux positifs et des faux négatifs subsistent pour plusieurs classes. Dans les vidéos réelles, ces erreurs sont majoritairement dues aux occlusions partielles entre les boxeurs, aux changements rapides de point de vue, au flou de mouvement ainsi qu'aux erreurs d'estimation de pose provoquées par des postures extrêmes ou des contacts rapprochés. Ces perturbations affectent directement la qualité des squelettes extraits, qui constituent l'entrée principale du modèle. Du côté des données synthétiques, bien que les mouvements soient mieux structurés, un écart de domaine persiste entre les scènes générées artificiellement et les scènes filmées. Cet écart peut entraîner des erreurs de généralisation lorsque le modèle est confronté à des configurations visuelles réelles qu'il n'a jamais observées durant l'entraînement.

Par ailleurs, la proximité gestuelle entre certaines classes, notamment entre les jabs, les feintes et les déplacements courts, favorise des confusions inter-classes, en particulier lors des transitions rapides. L'influence potentielle des couleurs des vêtements sur le suivi d'identité et la discrimination des boxeurs constitue un point intéressant à discuter. Dans certains cas, des contrastes chromatiques marqués entre les boxeurs peuvent faciliter le suivi visuel et la séparation des trajectoires, ce qui peut indirectement améliorer la stabilité du tracking et la cohérence des séquences de poses utilisées pour la classification comme par exemple les couleurs des tenues sont toujours rouges et bleus lors des combats olympiques. Toutefois, cette approche ne constitue pas une solution générale, car elle devient rapidement inefficace en présence d'occlusions, de variations d'éclairage ou de changements de caméra importants. Elle doit donc être considérée comme un facteur auxiliaire, susceptible d'être intégré dans un cadre plus large de randomisation de domaine ou d'adaptation de domaine mais ne saurait à elle seule garantir une robustesse suffisante dans des conditions réelles variées.

Une autre limite importante réside dans l'absence d'évaluation explicite de la qualité de l'estimation de pose sur les données synthétiques. Contrairement aux vidéos réelles, pour lesquelles la pose est estimée à partir d'images bruitées, la génération synthétique permet théoriquement d'accéder à une vérité terrain complète en trois dimensions. La comparaison entre les poses 2D extraites à partir des rendus et les poses 3D exactes issues du moteur de rendu permettrait de quantifier précisément l'erreur d'estimation et d'isoler la part d'erreur imputable à la détection de pose de celle liée à la classification des actions. L'absence de cette analyse

limite, à ce stade, l'interprétation fine des performances obtenues.

Enfin, il convient de souligner que le modèle a été entraîné sur un sous-ensemble annoté du jeu de données de base de 1536 vidéos synthétiques du corpus total qui son non-intégration dans l'apprentissage supervisé limite l'impact maximal du synthétique sur les performances du modèle. L'extension future de l'annotation à une part plus large de ce corpus pourrait conduire à des gains supplémentaires significatifs. En perspective, plusieurs axes d'amélioration se dégagent ici. Ce corpus non annoté ouvre également la voie à des approches non supervisées et auto-supervisées pour l'apprentissage des représentations spatiotemporelles. Un tel cadre permettrait d'exploiter pleinement le potentiel du volume de données générées sans dépendre exclusivement du coût d'annotation. Par ailleurs, l'enrichissement de la variabilité visuelle des scènes synthétiques (textures, vêtements, éclairages dynamiques, bruit visuel, occlusions artificielles) ainsi que l'intégration de méthodes de randomisation ou d'adaptation de domaine contribueraient à réduire l'écart de distribution entre les données réelle et synthétique. Enfin, l'introduction d'une évaluation quantitative de l'erreur de pose, rendue possible par l'accès à la vérité terrain 3D, permettrait d'analyser plus finement les sources d'erreurs du pipeline global et d'orienter plus précisément les améliorations futures, tant en apprentissage supervisé qu'en apprentissage non supervisé. Cette discussion met ainsi en évidence que, bien que l'approche hybride réelle et synthétique constitue une solution efficace pour la reconnaissance d'actions en boxe et ouvre la voie à des travaux futurs visant à renforcer encore la fiabilité, la généralisabilité et la précision du système proposé.

4.5.6 Effet des Données Synthétiques sur les Performances

L'un des objectifs majeurs de cette thèse est d'évaluer dans quelle mesure l'utilisation de données synthétiques peut améliorer la reconnaissance d'actions de boxe, en particulier dans un contexte où les données réelles sont rares, coûteuses à annoter et souvent bruitées (occlusions, variations de caméras, angles de vue limités). Les résultats obtenus démontrent que les données synthétiques jouent un rôle déterminant sur plusieurs aspects :

1. Amélioration des classes rares : les actions telles que les uppercuts ou les crochets arrière, sous-représentées dans les bases réelles, bénéficient significativement des séquences générées artificiellement. Le gain observé est de l'ordre de 15 points F1 pour l'uppercut.
2. Couverture étendue des comportements : la génération procédurale nous a permis d'inclure des classes souvent ignorées (feintes, esquives, déplacements), enrichissant la granularité de l'analyse.
3. Homogénéité des performances : contrairement au modèle de Seminario dont les résultats varient fortement selon les classes (de 48 à 86), notre modèle présente une plage plus resserrée (71 à 92), témoignant d'une meilleure généralisation.
4. Potentiel de complémentarité : si les données réelles restent essentielles pour capter la variabilité et le réalisme, les données synthétiques apparaissent comme un complément indispensable pour équilibrer les distributions de classes et renforcer la robustesse des modèles.

En conclusion, l'intégration des données synthétiques constitue une réponse efficace aux limitations structurelles rencontrées dans l'apprentissage supervisé à partir de vidéos réelles de boxe. Elle permet non seulement d'améliorer les performances sur les classes rares et difficiles, mais aussi d'élargir la couverture fonctionnelle des modèles vers une analyse plus complète et fidèle du sport.

CONCLUSION ET RECOMMANDATIONS

Cette thèse a exploré une approche innovante pour l'analyse automatique des performances en boxe, fondée sur la génération de données synthétiques de haute qualité et le développement d'un pipeline intégré d'analyse. Le travail réalisé couvre l'ensemble de la chaîne de traitement, depuis la génération et l'annotation des données jusqu'au suivi et à la classification des actions de boxe. L'ensemble du système proposé repose actuellement sur des scénarios comprenant exclusivement deux boxeurs interagissant dans le ring, ce qui a permis de concevoir un cadre d'analyse cohérent pour la détection et la reconnaissance des actions.

Dans la perspective de travaux futurs, une extension naturelle de ce cadre consisterait à intégrer la présence d'un arbitre dans la scène. Cette évolution nécessitera la collecte et l'annotation de nouvelles données spécifiques à la posture, au positionnement et aux déplacements de l'arbitre, afin de permettre l'apprentissage d'un modèle capable de le détecter et de le distinguer des boxeurs pendant les séquences. Une telle intégration renforcerait la robustesse du système d'analyse, en permettant une interprétation plus complète des interactions dans les vidéos de combat. Au-delà de cette extension, les perspectives à long terme de ce travail s'inscrivent dans le développement de systèmes d'analyse multi-agents capables de modéliser de manière fine la dynamique spatio-temporelle des combats. Une telle approche ouvrirait la voie à une compréhension plus riche des performances sportives, contribuant à l'avancement des technologies d'intelligence artificielle appliquées à l'analyse du mouvement humain et à la science du sport.

ANNEXE I

GLOSSAIRE

Ce glossaire regroupe l'ensemble des termes, abréviations et sigles techniques utilisés dans ce mémoire accompagnés de leurs définitions. Il vise à faciliter la lecture et la compréhension des notions employées dans le cadre de ce travail de recherche.

AI (Artificial Intelligence) Domaine de l'informatique visant à créer des systèmes capables d'effectuer des tâches nécessitant l'intelligence humaine, telles que la reconnaissance, la planification ou la prise de décision.

ANN (Artificial Neural Network) Modèle d'apprentissage inspiré du fonctionnement des neurones biologiques, utilisé pour reconnaître des motifs complexes dans les données.

Annotation Attribution de labels aux vidéos ou images pour identifier des actions, des postures ou des résultats. Peut être manuelle (faite par des experts), automatique (via simulation) ou semi-automatique (combinaison des deux).

AP (Average Precision) Mesure d'évaluation utilisée en détection d'objets et en reconnaissance d'actions pour combiner précision et rappel en une seule métrique.

Blender Logiciel libre de modélisation, d'animation et de rendu 3D. Utilisé ici comme moteur de génération d'environnements virtuels et de vidéos synthétiques de boxe.

BVH (BioVision Hierarchy) Format standard ancien pour représenter les mouvements humains. Il décrit un squelette hiérarchique dans lequel chaque articulation est définie par ses rotations au fil du temps. Dans ce mémoire, le BVH constitue une référence historique, bien qu'il ait été supplanté par des formats plus complets comme le FBX.

CG (Computer Graphics) Domaine de l'informatique dédié à la création, la manipulation et le rendu d'images et d'environnements 3D.

CNN (Convolutional Neural Network) Architecture de réseau de neurones profonde utilisée pour l'analyse d'images, particulièrement efficace pour la détection et la reconnaissance visuelle.

COCO (Common Objects in Context) Jeu de données de référence en vision par ordinateur, utilisé pour l'entraînement et l'évaluation de modèles de détection d'objets et de poses.

CSV (Comma-Separated Values) Format de fichier tabulaire utilisé pour stocker des annotations ou des séries de données numériques de manière structurée.

Jeu de données synthétique Jeu de données généré artificiellement à l'aide de modèles 3D, de moteurs de rendu et de la capture de mouvement. Il offre des avantages tels que la diversité des morphologies, l'équilibre des classes et la génération automatique d'annotations. Utilisé ici pour pallier la rareté des vidéos de boxe annotées.

DL (Deep Learning) Sous-domaine de l'intelligence artificielle basé sur l'entraînement de réseaux de neurones profonds pour l'apprentissage de représentations hiérarchiques à partir de données.

DOF (Degrees of Freedom / Degrés de Liberté) Nombre de variables indépendantes nécessaires pour décrire la configuration d'un système articulé. Exemple : le coude = 1 DOF, l'épaule = 3 DOF. Utilisé ici pour relier la représentation numérique à la réalité biomécanique, élément essentiel pour caractériser la complexité des gestes de boxe.

Estimation de pose Processus visant à localiser les articulations clés d'un corps humain dans une image ou une vidéo. L'estimation 2D fournit les coordonnées (x, y), tandis que l'estimation 3D inclut la profondeur (z). Des modèles modernes comme OpenPose, HRNet ou ViTPose exploitent des réseaux de neurones profonds et des Transformers.

FBX (FilmBoX) Format développé par Autodesk et devenu la norme dans les domaines de l'animation et de la vision par ordinateur. Il permet de stocker dans un seul fichier les géométries 3D, les textures, les matériaux et les animations. Dans ce travail, le FBX est utilisé comme format pivot pour intégrer les mouvements de boxeurs dans Blender, appliquer SMPL et générer des vidéos synthétiques réalistes.

FPS (Frames Per Second) Unité de mesure du nombre d'images affichées ou rendues par seconde, utilisée pour évaluer la fluidité d'une vidéo ou d'un rendu.

GCN (Graph Convolutional Network) Type de réseau de neurones appliquant des convolutions sur des graphes, utilisé pour modéliser les relations entre articulations dans les squelettes humains.

- GPU (Graphics Processing Unit)** Processeur spécialisé dans les calculs parallèles, essentiel pour l'entraînement de modèles de deep learning et le rendu 3D.
- HOG (Histogram of Oriented Gradients)** Descripteur visuel classique utilisé pour détecter des formes et des silhouettes à partir de gradients d'intensité dans les images.
- HOF (Histogram of Optical Flow)** Descripteur spatio-temporel représentant les mouvements dans une séquence vidéo à partir des flux optiques locaux.
- HPJRE (High-Precision Joint Reconstruction Error)** Mesure de précision pour évaluer l'erreur de reconstruction des articulations dans les modèles de pose 3D.
- HRNet (High-Resolution Network)** Architecture de réseau conçue pour maintenir des représentations à haute résolution tout au long du processus d'estimation de pose.
- I3D (Inflated 3D ConvNet)** Réseau convolutif 3D dérivé de modèles 2D, utilisé pour la reconnaissance d'actions dans les vidéos.
- INS Québec** Institut National du Sport du Québec, partenaire scientifique et sportif de ce projet.
- LSP (Leeds Sports Pose)** Jeu de données public contenant des poses humaines annotées, utilisé comme référence pour l'évaluation de modèles d'estimation de pose.
- LSF (Long Short-Term Feature)** Représentation temporelle utilisée pour capturer les dépendances séquentielles dans les données de mouvement.
- LSTM (Long Short-Term Memory)** Type de réseau récurrent conçu pour traiter les séquences temporelles et éviter la perte de mémoire à long terme.
- ML (Machine Learning)** Branche de l'intelligence artificielle reposant sur des algorithmes capables d'apprendre automatiquement à partir de données.
- Métriques (Précision, rappel, F1-score)** Mesures standard utilisées pour évaluer les performances des modèles de détection de pose et de classification d'actions.
- MiXSTE (Mixed Spatio-Temporal Encoder)** Modèle de deep learning combinant des encodeurs spatiaux et temporels pour analyser des séquences d'actions.
- MPJPE (Mean Per Joint Position Error)** Indicateur d'évaluation de la précision de l'estimation de pose 3D, calculé comme l'erreur moyenne par articulation.

MPII (Max Planck Institute for Informatics) Jeu de données de référence pour l'estimation de poses humaines en 2D.

Motion Capture (MoCap) Technologie de capture de mouvement permettant d'enregistrer les gestes réels des boxeurs et de les réappliquer à des modèles virtuels.

Occlusion Situation dans laquelle une partie du corps est masquée dans une vidéo (par l'adversaire, les gants ou l'angle de caméra). Ce phénomène constitue un défi majeur pour la détection et l'estimation de pose.

OKS (Object Keypoint Similarity) Mesure de similarité utilisée pour évaluer la précision de la détection de points clés dans les images.

OpenPose Framework open source d'estimation de pose humaine, basé sur les champs d'affinité de parties (PAFs).

PAFs (Part Affinity Fields) Représentations vectorielles reliant les articulations entre elles, utilisées pour associer les points clés appartenant à un même individu.

PCK (Percentage of Correct Keypoints) Taux de points clés correctement détectés dans une tolérance spatiale donnée.

ProtoGCN Prototype Graph Convolutional Network – architecture basée sur des graphes qui exploite les relations articulaires pour reconnaître des actions humaines.

ReLU (Rectified Linear Unit) Fonction d'activation utilisée dans les réseaux de neurones pour introduire la non-linéarité tout en réduisant le problème du gradient nul.

Reconnaissance d'actions spatio-temporelles Classification automatique d'actions (jab, crochet, esquive, blocage, etc.) à partir de séquences vidéo. Les approches vont des descripteurs manuels (STIP, HOG, HOF) aux architectures profondes modernes telles que I3D, SlowFast, ST-GCN ou ProtoGCN.

Retargeting Technique permettant de transférer un mouvement capturé d'un individu vers un modèle virtuel d'une morphologie différente, tout en préservant la fidélité biomécanique du geste.

R-MPJPE (Root Mean Per Joint Position Error) Variante du MPJPE calculée après alignement du squelette sur le bassin, pour évaluer la précision relative des poses.

RNN / LSTM (Recurrent Neural Networks / Long Short-Term Memory) Réseaux récurrents utilisés pour traiter des séquences temporelles. En analyse de la boxe, ils permettent de modéliser l'évolution dynamique des mouvements.

Réseaux de neurones convolutionnels (CNN) Modèles d'apprentissage profond conçus pour l'analyse d'images. En boxe, ils sont utilisés pour détecter les boxeurs, estimer les poses et extraire des caractéristiques visuelles discriminantes.

SMPL (Skinned Multi-Person Linear Model) Modèle statistique du corps humain reposant sur un maillage triangulaire déformable, contrôlé par des paramètres de forme et de pose. Les extensions SMPL-X (ajout des mains et du visage) et STAR (prise en compte des muscles) permettent d'obtenir une représentation plus complète.

SMPL-X Extension du modèle SMPL intégrant la modélisation des mains, du visage et de la posture complète.

ST-GCN (Spatio-Temporal Graph Convolutional Network) Réseau exploitant la structure squelettique sous forme de graphe (nœuds = articulations, arêtes = connexions anatomiques). Appliqué à la reconnaissance d'actions à partir des coordonnées articulaires.

STIP (Space-Time Interest Points) Méthode classique de détection de points d'intérêt spatio-temporels pour l'analyse des actions dans les vidéos.

TCP (Temporal Convolutional Pooling) Mécanisme de convolution temporelle utilisé pour agréger les informations sur plusieurs trames consécutives.

Transformers Architectures d'apprentissage profond reposant sur des mécanismes d'attention, capables de capturer des dépendances spatiales et temporelles longues. Adaptées à la vision (ViT, ViTPose) et à la vidéo (Video Transformer).

UE (Unreal Engine) Moteur 3D temps réel utilisé pour la modélisation, la simulation et le rendu d'environnements virtuels interactifs.

ViTPose Vision Transformer for Pose Estimation – modèle basé sur les Transformers pour estimer les articulations humaines avec haute précision.

YOLO (You Only Look Once) Algorithme de détection d'objets en temps réel capable de localiser et classer plusieurs objets dans une seule passe de réseau.

ANNEXE II

ACTIONS DE BOXE EN DESCRIPTION

ID	Action	Description
1	Jab	Coup direct rapide du bras avant, visant à contrôler la distance et perturber l'adversaire.
2	Cross	Coup direct puissant du bras arrière, généralement utilisé comme coup final ou en combinaison.
3	Lead Hook	Crochet circulaire du bras avant, visant la tête ou le corps de l'adversaire.
4	Rear Hook	Crochet circulaire du bras arrière, plus puissant, ciblant la tête ou le corps.
5	Lead Uppercut	Uppercut montant avec le bras avant, utilisé à courte distance pour surprendre l'adversaire.
6	Rear Uppercut	Uppercut montant du bras arrière, souvent utilisé pour casser une garde rapprochée.
7	Rear Overhand	Coup descendant et circulaire du bras arrière, visant la tête par-dessus la garde adverse.
8	Neutral	Position de garde de base, sans action offensive ou défensive spécifique.
9	Slip	Déplacement latéral de la tête pour esquiver un coup direct de l'adversaire.
10	Bob and Weave	Mouvement d'évitement combinant flexion et rotation du tronc pour passer sous un coup.
11	Pull	Retrait rapide du buste vers l'arrière pour esquiver un coup.
12	Duck	Abaissement brusque en fléchissant les genoux pour esquiver un coup haut.
13	Parry	Déviation d'un coup adverse avec la main, redirigeant sa trajectoire.
14	Block	Blocage d'un coup adverse en utilisant la garde ou les bras.
15	Clinch	Corps-à-corps défensif pour neutraliser les attaques rapprochées.
16	Pivot	Rotation sur un appui pour changer d'angle et créer une ouverture.
17	Angle off	Déplacement latéral rapide pour sortir de l'axe d'attaque et se repositionner.
18	Punch Feint	Simulation d'un coup pour tromper la défense de l'adversaire.
19	Shoulder Feint	Mouvement d'épaule simulant le déclenchement d'une attaque.
20	Feet / step feint	Simulation de déplacement pour désorienter la défense adverse.

Tableau-A II-1 Détails des actions boxe pour annotation

ANNEXE III

EXEMPLES D' ACTIONS DE BOXE EN IMAGES



Figure-A III-1 Crochet de la main avant à la tête



Figure-A III-2 Crochet de la main arrière à la tête

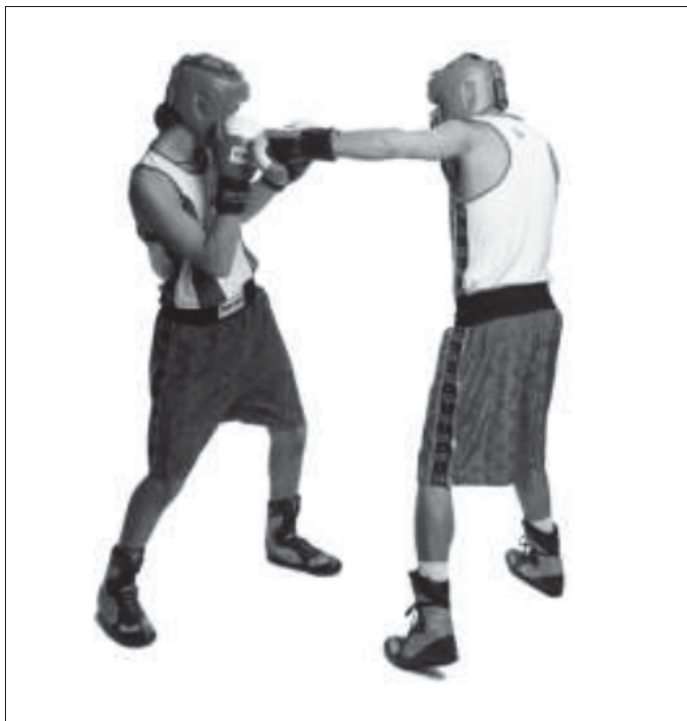


Figure-A III-3 Bloc à l'aide de l'avant-bras

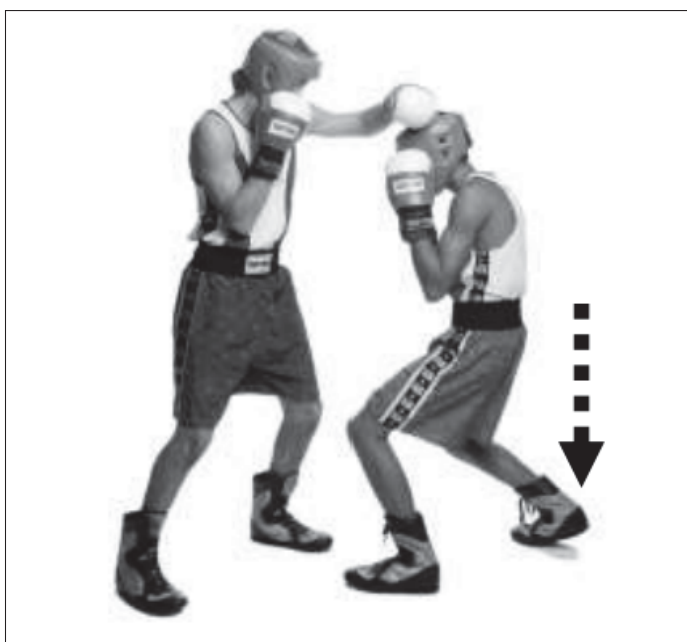


Figure-A III-4 Évitement vers le bas



Figure-A III-5 Jab, pas avant



Figure-A III-6 Direct de la main arrière à la tête



Figure-A III-7 Direct de la main arrière au corps

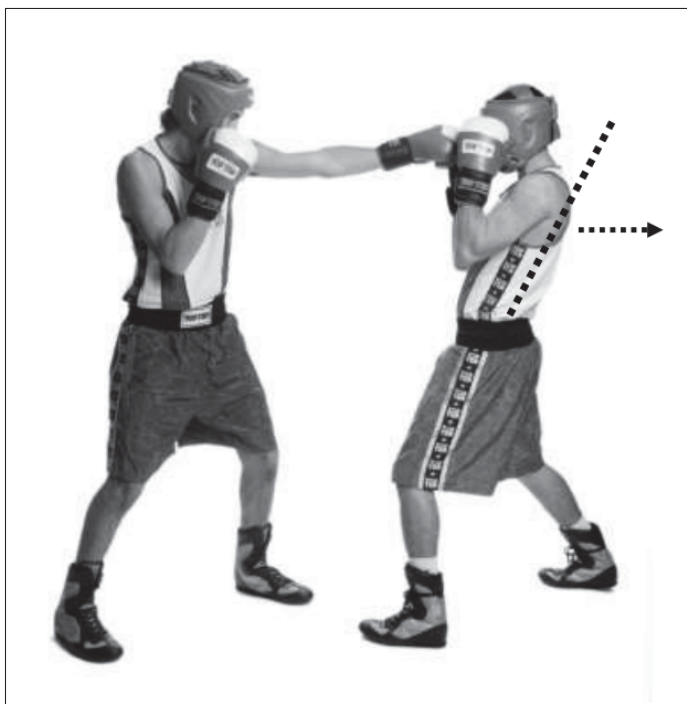


Figure-A III-8 Esquive vers l'arrière



Figure-A III-9 Uppercut au corps



Figure-A III-10 Jab contre un crochet de la main avant dirigé au corps

ANNEXE IV

LES PIPELINES DE LA CRÉATION DE JEU DE DONNÉES SYNTHÉTIQUE

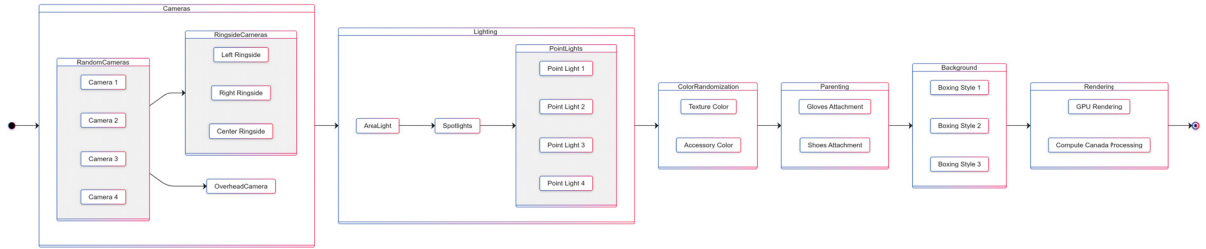


Figure-A IV-1 Pipeline de création du ring de boxe

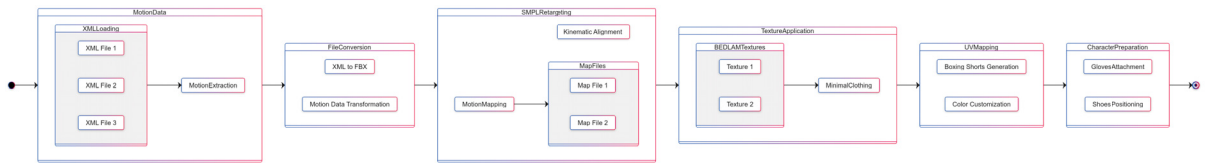


Figure-A IV-2 Pipeline de la création du boxeur

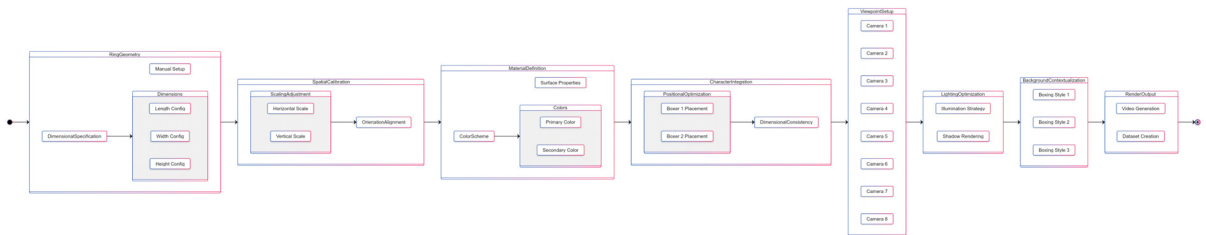


Figure-A IV-3 Pipeline du rendu des scènes synthétiques

ANNEXE V

EXEMPLES DE CLASSIFICATION DES ACTIONS DE BOXE EN IMAGES

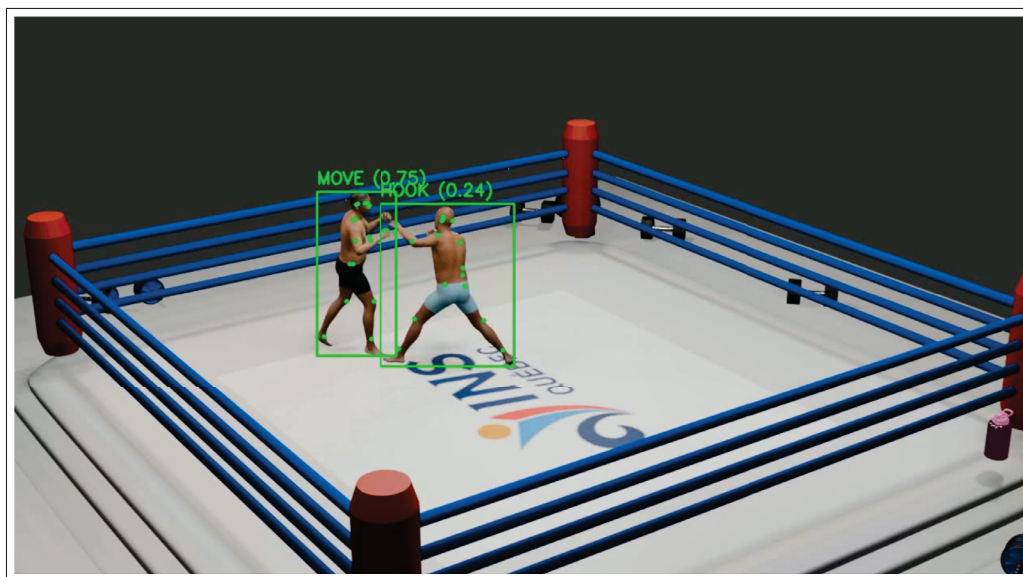


Figure-A V-1 MoveVsHook

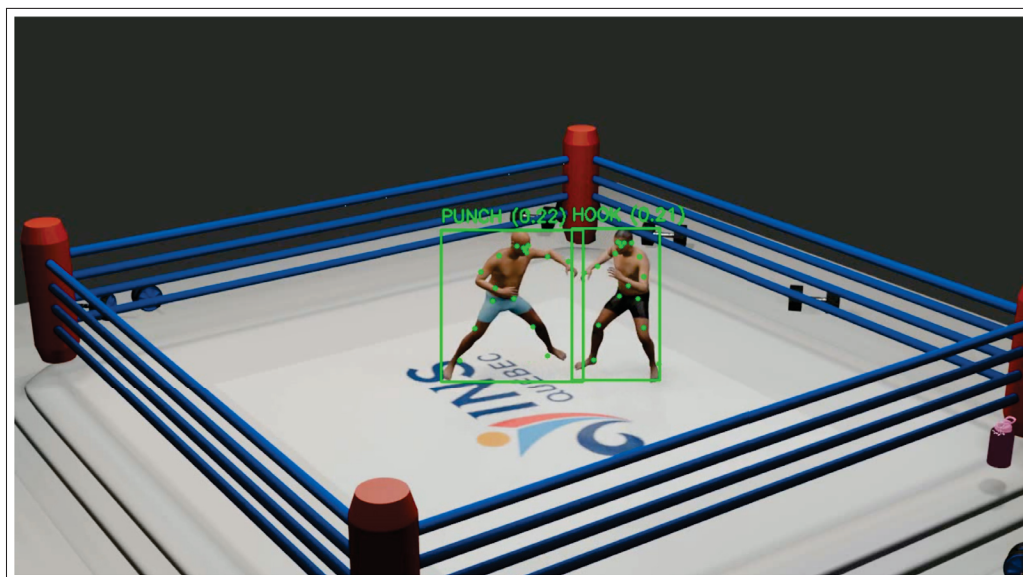


Figure-A V-2 PunchVsHook

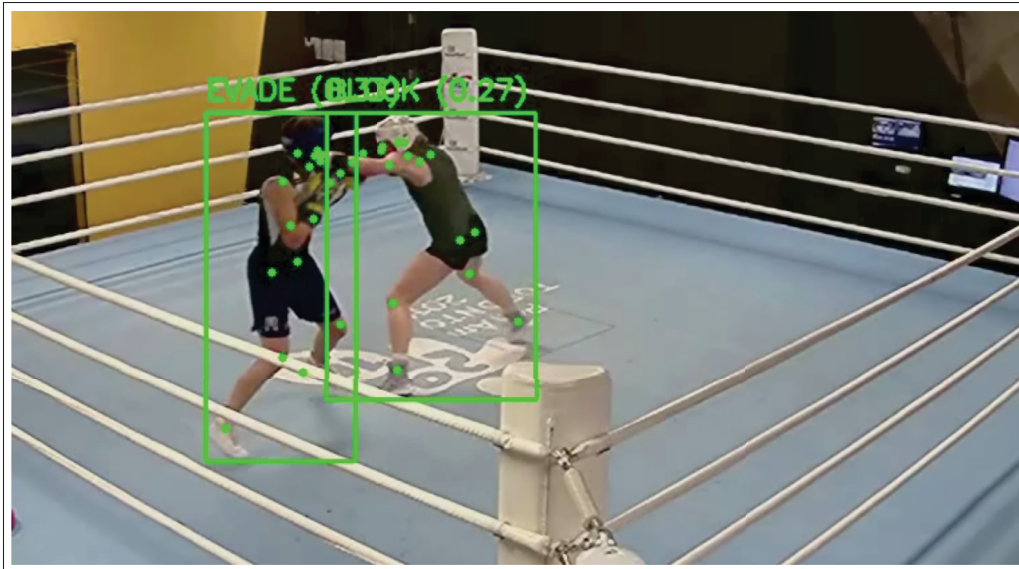


Figure-A V-3 EvadeVsHook

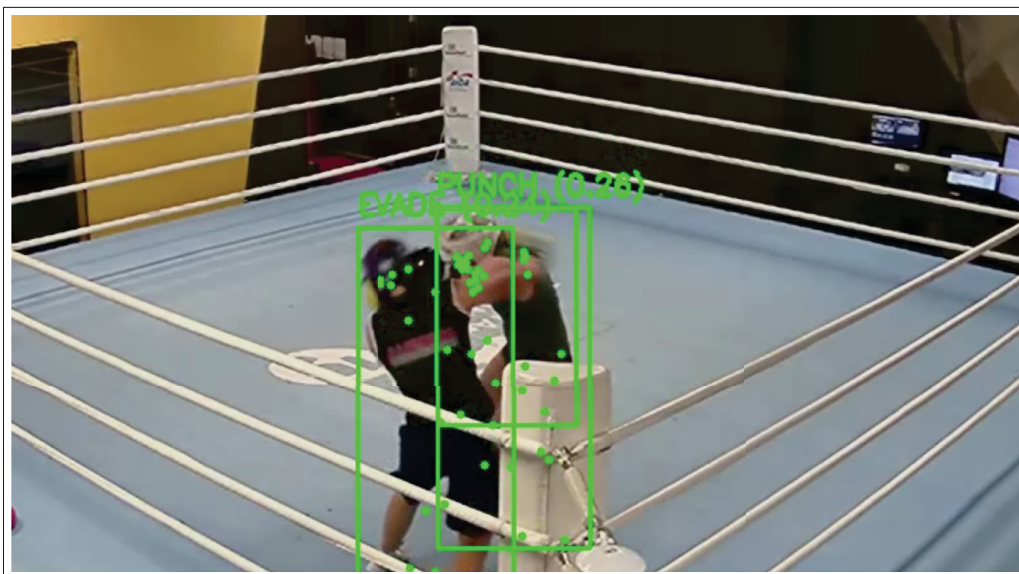


Figure-A V-4 PunchVsEvade

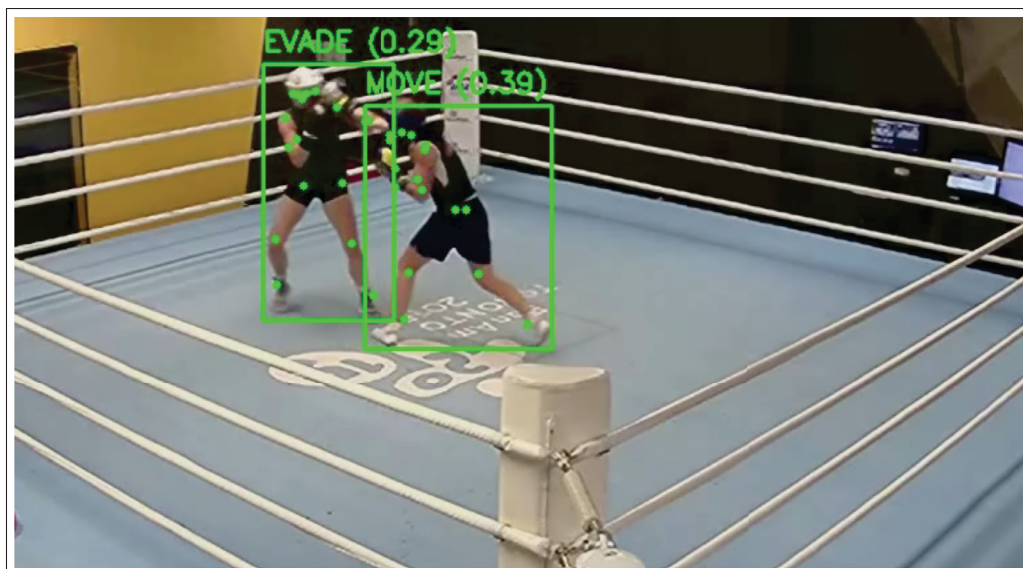


Figure-A V-5 MoveVsEvade

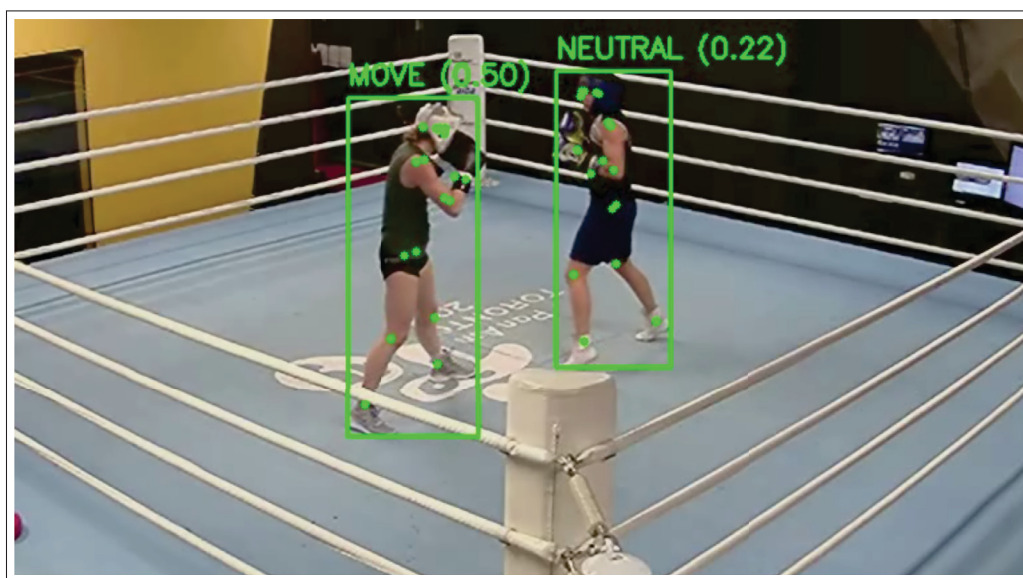


Figure-A V-6 EMoveVsNeutral

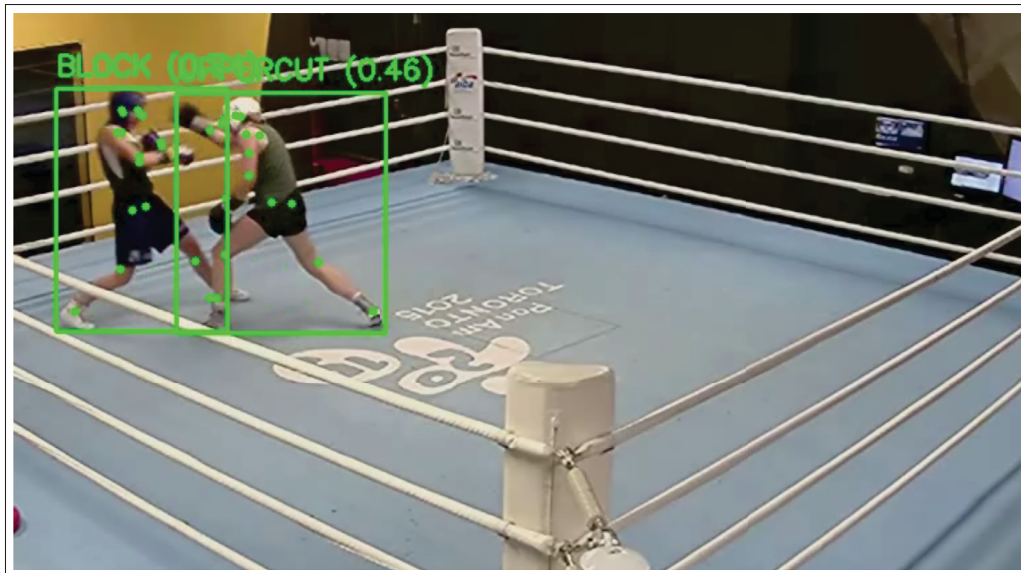


Figure-A V-7 UppercutVsBlock

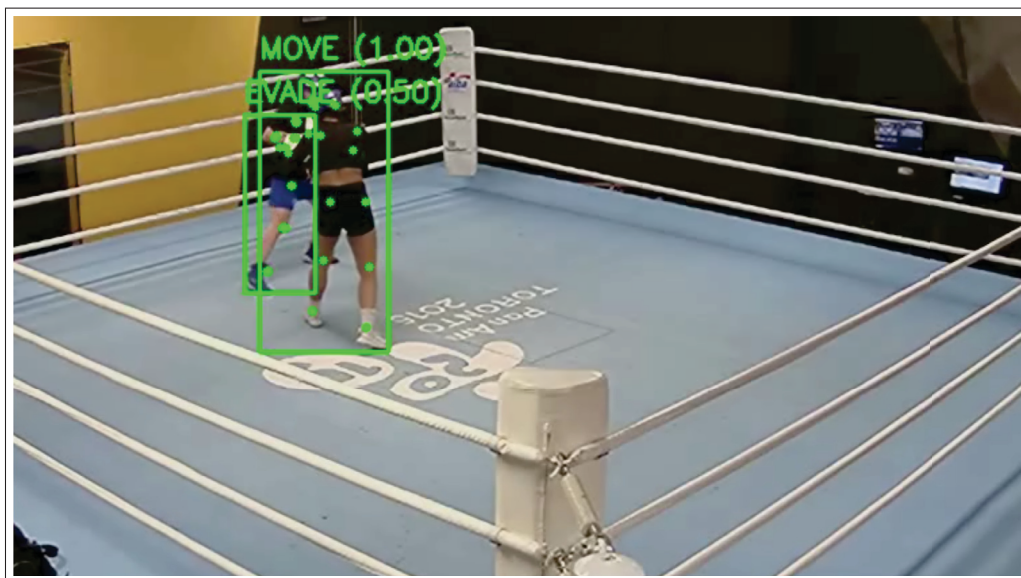


Figure-A V-8 EvadeVsMove

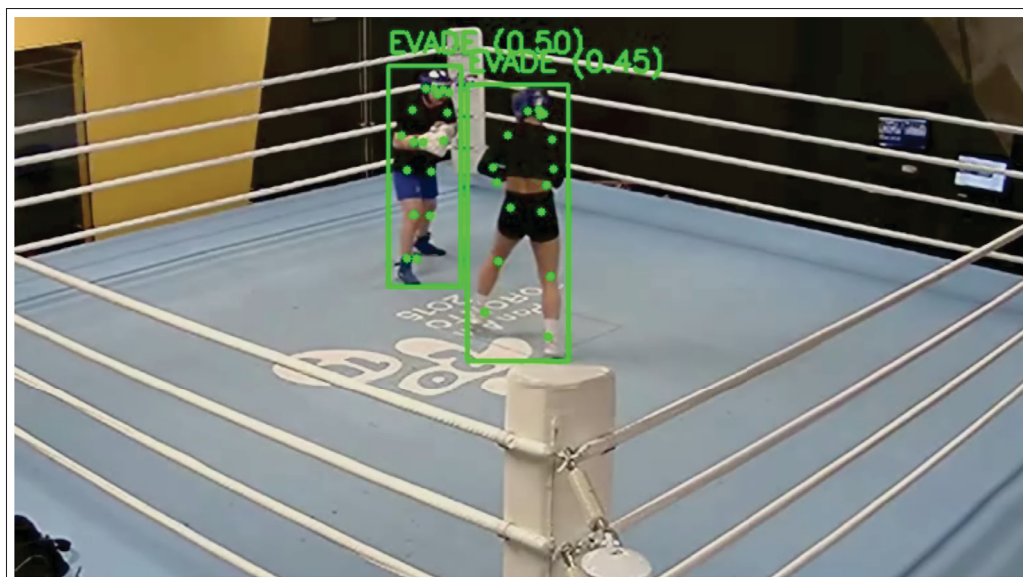


Figure-A V-9 EvadeVsEvade

BIBLIOGRAPHIE

- Aberman, K., Li, P., Lischinski, D., Sorkine-Hornung, O., Cohen-Or, D. & Chen, B. (2020). Skeleton-aware networks for deep motion retargeting. *ACM Transactions on Graphics (TOG)*, 39(4), 62–1.
- Aristidou, A., Lasenby, J., Chrysanthou, Y. & Shamir, A. (2018). Inverse kinematics techniques in computer graphics : A survey. *Computer graphics forum*, 37(6), 35–58.
- Bartlett, R. (2014). *Introduction to sports biomechanics : Analysing human movement patterns*. Routledge.
- Black, M. J., Patel, P., Tesch, J. & Yang, J. (2023). Bedlam : A synthetic dataset of bodies exhibiting detailed lifelike animated motion. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8726–8737.
- Cao, Z., Hidalgo, T. S., Wei, S.-E. & Sheikh, Y. (2019). OpenPose : Realtime multi-person 2D pose estimation using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.
- Carreira, J. & Zisserman, A. (2017). Quo vadis, action recognition ? A new model and the kinetics dataset. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6299–6308.
- Chen, D., Song, Y., Liang, F., Ma, T., Zhu, X. & Jia, T. (2023). 3D human body reconstruction based on SMPL model. *The Visual Computer*, 39(5), 1893–1906.
- Dalal, N., Triggs, B. & Schmid, C. (2006). Human detection using oriented histograms of flow and appearance. *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 428–441.
- Das Dawn, D. & Shaikh, S. H. (2016). A comprehensive survey of human action recognition with spatio-temporal interest point (STIP) detector. *The Visual Computer*, 32, 289–306.
- Dong, C. & Du, G. (2024). An enhanced real-time human pose estimation method based on modified YOLOv8 framework. *Scientific Reports*, 14(1), 8012.
- Feichtenhofer, C., Fan, H., Malik, J. & He, K. (2019). SlowFast networks for video recognition. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6202–6211.

- Gao, Z., Chen, J., Liu, Y., Jin, Y. & Tian, D. (2025). A systematic survey on human pose estimation : upstream and downstream tasks, approaches, lightweight models, and prospects. *Artificial Intelligence Review*, 58(3), 68.
- Gleicher, M. (1998). Retargetting motion to new characters. *Proceedings of the 25th annual conference on Computer graphics and interactive techniques (SIGGRAPH)*, pp. 33–42.
- Guo, X., Wu, W., Wang, D., Su, J., Su, H., Gan, W., Huang, J. & Yang, Q. (2022). Learning video representations of human motion from synthetic data. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20197–20207.
- Kasiri, S., Fookes, C., Sridharan, S. & Morgan, S. (2017). Fine-grained action recognition of boxing punches from depth imagery. *Computer Vision and Image Understanding*, 159, 143–153.
- Khoiee, H. F. Z., Labbe, D., Romeas, T., Faubert, J. & Andrews, S. (2025). Multi-person Physics-based Pose Estimation for Combat Sports. *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5832–5841.
- Laptev, I. (2005). Space-time interest points. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 432–439.
- Liu, H., Liu, Y., Ren, M., Wang, H., Wang, Y. & Sun, Z. (2025). Revealing key details to see differences : A novel prototypical perspective for skeleton-based action recognition. *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29248–29257.
- Loper, M., Mahmood, N., Romero, J., Pons-Moll, G. & Black, M. J. (2015). SMPL : A skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6), 1–16.
- Martinez, J., Hossain, R., Romero, J. & Little, J. J. (2017). A simple yet effective baseline for 3d human pose estimation. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 2640–2649.
- Newell, A., Yang, K. & Deng, J. (2016). Stacked hourglass networks for human pose estimation. *Computer Vision–ECCV 2016 : 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VIII 14*, pp. 483–499.
- Osman, A. A. A., Bolkart, T. & Black, M. J. (2020). STAR : Sparse trained articulated human body regressor. *European Conference on Computer Vision (ECCV)*, pp. 598–613.

- Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A. A. A., Tzionas, D. & Black, M. J. (2019). Expressive body capture : 3D hands, face, and body from a single image. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10975–10985.
- Pavlo, D., Feichtenhofer, C., Grangier, D. & Auli, M. (2019). 3D human pose estimation in video with temporal convolutions and semi-supervised training. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7753–7762.
- Prakash, A., Boochoon, S., Brophy, M., Acuna, D., Cameracci, E., State, G., Shapira, O. & Birchfield, S. (2019). Structured domain randomization : Bridging the reality gap by context-aware synthetic data. *2019 International Conference on Robotics and Automation (ICRA)*, pp. 7249–7255.
- Roberto de Souza, C., Gaidon, A., Cabon, Y. & Manuel Lopez, A. (2017). Procedural generation of videos to train deep action recognition networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4757–4767.
- Rokoko. [Accessed : Oct. 2025]. (2023). Blender Integration & Inverse Kinematics Guide. Repéré à <https://help.rokoko.com/hc/en-us/articles/4580996884764-Blender-integration>.
- Seminaro, J. (2022). *Développement d'une méthode d'extraction automatique de métriques de performance en boxe basée sur la vision par ordinateur*. (Mémoire de maîtrise, École de technologie supérieure, Montréal).
- Sengupta, S., Budvytis, I. & Cipolla, R. (2020). Synthetic training for accurate 3D human pose and shape estimation in the wild. *British Machine Vision Conference (BMVC)*.
- Sohan, M., Sai Ram, T. & Rami Reddy, C. V. (2024). A review on yolov8 and its advancements. *International Conference on Data Intelligence and Cognitive Informatics*, pp. 529–545.
- Stefański, P., Kozak, J. & Jach, T. (2024). Boxing Punch Detection with Single Static Camera. *Entropy*, 26(8), 617.
- Sun, K., Xiao, B., Liu, D. & Wang, J. (2019). High-resolution representations for labeling pixels and regions. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2454–2463.
- Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W. & Abbeel, P. (2017). Domain randomization for transferring deep neural networks from simulation to the real world. *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pp. 23–30.

- Toshev, A. & Szegedy, C. (2014). Deeppose : Human pose estimation via deep neural networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1653–1660.
- Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M. J., Laptev, I. & Schmid, C. (2017). Learning from synthetic humans. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 109–117.
- Veeriah, V., Zhuang, N. & Qi, G.-J. (2015). Differential recurrent neural networks for action recognition. *Proceedings of the IEEE international conference on computer vision*, pp. 4041–4049.
- Villegas, R., Yang, J., Zou, Y., Sohn, K., Lin, X. & Lee, H. (2018). Learning to generate long-term future via hierarchical prediction. *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 3560–3569.
- Wang, X., Miao, Z., Zhang, R. & Hao, S. (2019a). I3d-lstm : A new model for human action recognition. *IOP conference series : materials science and engineering*, 569(3), 032035.
- Wang, X., Li, L., Ye, W., Long, M. & Wang, J. (2019b). Transferable attention for domain adaptation. *Proceedings of the AAAI conference on artificial intelligence*, 33(01), 5345–5352.
- Xiao, B., Wu, H. & Wei, Y. (2018). Simple baselines for human pose estimation and tracking. *Proceedings of the European conference on computer vision (ECCV)*, pp. 466–481.
- Xu, Y., Zhang, Z., Xu, K., Geng, Z., Wang, T., Yang, L., Wang, L. & Wei, Y. (2022). ViTPose : Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems (NeurIPS)*.
- Yan, S., Xiong, Y. & Lin, D. (2018). Spatial temporal graph convolutional networks for skeleton-based action recognition. *Proceedings of the AAAI conference on artificial intelligence*, 32(1).
- Zhang, J., Tu, Z., Yang, J., Chen, Y. & Yuan, J. (2022). Mixste : Seq2seq mixed spatio-temporal encoder for 3d human pose estimation in video. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13232–13242.
- Zhao, Z. & Chen, N.-Z. (2023). Spatial-temporal graph convolutional networks (STGCN) based method for localizing acoustic emission sources in composite panels. *Composite Structures*, 323, 117496.