الجمهورية العربية السورية

المعهد العالي للعلوم التطبيقية والتكنولوجيا

قسم المعلوميات

تطوير نظام لإعادة التعرف من خلال دمج عدة خصائص بيومترية في بيئة معطيات كبيرة

Developing A Person Re-ID System through Fusing Multiple Biometrics in a Big Data Environment

أطروحة بحث لرسالة في ماجستير معلوماتية اختصاص نظم المعطيات الكبيرة

تقديم: م. عبد الرحمن أحمد الأحمر

إشراف: د. مصطفى دقاق

د. باسم السهوة

لجنة الحكم

د.عمر حمدون المعهد العالي للعلوم التطبيقية والتكنولوجيا

رئيساً ومقرراً

د.ميساء أبو قاسم جامعة دمشق

مقرراً

د.رياض سنبل المعهد العالي للعلوم والتطبيقية والتكنولوجيا

مقرراً

د.مصطفى دقاق المعهد العالي للعلوم التطبيقية والتكنولوجيا

مشرفاً ومقرراً

تصريح

أنا الموقع أدناه **معدّ أطروحة الماجستير التي تحمل العنوان:**

أصرح بأن:

– الأعمال والنتائج المعروضة في هذه الأطروحة هي نتيجة جهودي الشخصية وبتوجيه من المشرف، وأن ما عدا ذلك من معلومات ونتائج قد نُسبت إلى مصادرها ومؤلفيها، وأُشير إلى ذلك في متن النص وفي قائمة المراجع.

– المعطيات والمعلومات المستخدمة في هذه الأطروحة جرى تحصيلها بطرائق سليمة ومشروعة ونُسبت إلى مصادرها في المواضع الملائمة.

– كلّ مكوّن من مكونات هذه الأطروحة (مقطع نصّي، صورة، مخطط، ...) مقتبس من عمل آخر جرى تمييزه بوضوح ونُسب إلى مصدره.

– الأعمال والنتائج المعروضة في هذه الأطروحة لم تُستخدم سابقًا وليست قيد الاستخدام للحصول على أي شهادة أكاديمية أخرى.

التوقيع

دمشق / / 202

المعهد العالي للعلوم التطبيقية والتكنولوجيا

Higher Institute for Applied Sciences and Technology

المعهد العالي للعلوم التطبيقية والتكنولوجيا مؤسسة حكومية للتعليم العالي أحدثت بموجب المرسوم التشريعي رقم /24/ لعام 1983، وذلك بهدف إعداد كوادر علمية متميزة من مهندسين وباحثين للإسهام الفاعل في عملية التطوير العلمي والتنمية في الجمهورية العربية السورية.

يمنح المعهد العالي درجة الإجازة في الهندسة في الاتصالات والمعلوماتية والنظم الإلكترونية والميكاترونيكس وعلوم وهندسة المواد وهندسة الطيران. يقبل المعهد العالي لدراسة هذه الاختصاصات شريحة منتقاة من المتفوقين في الشهادة الثانوية من الفرع العلمي. يتيح المعهد العالي أيضاً برامج ماجستير أكاديمي في نظم الاتصالات وفي التحكم والروبوتيك وفي نظم المعطيات الكبيرة ونظم المعلومات ودعم القرار وفي علوم وهندسة المواد وعلوم وهندسة البصريات. وأخيراً، يمنح المعهد العالي درجة الدكتوراه في الاتصالات والمعلوماتية ونظم التحكم والفيزياء التطبيقية. تُحدث في المعهد العالي اختصاصات جديدة بحسب متطلبات سوق العمل وتوجهات البحث والتطوير المحلية والعالمية.

إلى جانب النشاط التعليمي، يمارس المعهد العالي عبر جهود أطره وفعالياته العلمية المختلفة نشاطاً حثيثاً في البحث والتطوير، إذ ينفذ مشاريع ودراسات واستشارات متنوعة لصالح الجهات العامة والخاصة في القطر، كما يتعاون مع جهات خارج القطر في بعض المشاريع البحثية والتطويرية. يسعى المعهد أيضاً، عبر دورات تدريبية نظرية وعملية متاحة للقطاعين العام والخاص وللأفراد، إلى إفادة أوسع فئة من المهتمين من إمكانيات أطره العلمية ومختبراته. واستكمالاً لدوره الرائد في مجال التعليم ونشر العلم، ينشر المعهد العالي كتباً علمية عالية المستوى من نتاج أطره، منها ما هو تدريسي يوافق المناهج في المعهد العالي ويفيد شريحة واسعة من الطلاب الجامعيين عموماً، ومنها ما هو علمي ثقافي. يتيح المعهد العالي بعضاً من منشوراته على موقعه على الشابكة، كما يتيح إمكانية الاطلاع على رسائل الماجستير والدكتوراه المنفذة في المعهد العالي وعلى بعض منشورات طلابه وأطره من المقالات العلمية.

المعهد العالي للعلوم التطبيقية والتكنولوجيا، الجمهورية العربية السورية، دمشق، ص.ب 31983

Higher Institute for Applied Sciences and Technology – HIAST

P. O. Box 31983, Damascus, Syrian Arab Republic

هاتف 00963115123819 - فاكس 00963115140761

بريد إلكتروني contact@hiast.edu.sy

موقع إلكتروني www.hiast.edu.sy

الملخص

تعالج هذه الدراسة مسألة إعادة التعرّف على الأشخاص بوصفها مشكلة متعددة الأبعاد، تتداخل فيها الخصائص المظهرية، الزمنية، الحركية، والدلالية، ولا يمكن حصرها في تمثيل واحد أو مسار معالجة منفرد. هدفت الدراسة إلى تحليل نقدي لمقاربات إعادة التعرّف المعتمدة على المظهر، المشية، والوضعية، إضافة إلى المقاربات الحديثة التي توظّف النماذج اللغوية—البصرية لاستخلاص سمات دلالية عالية المستوى، مع إبراز نقاط القوة والقيود التشغيلية لكل فئة ضمن سياقات المراقبة الواقعية.

تشكل هذه الأطروحة مرجعاً غنياً وشاملاً لمسألة إعادة التعرف تغطي طرائق إعادة التعرف الشهيرة والمعمول بها في المسارات البحثية في السنوات الأخيرة بشكل واسع ومفصل.

اعتمدت المنهجية على دراسة مرجعية معمّقة مدعومة بتجارب عملية، ركّزت على تقييم فعالية استراتيجيات الدمج المختلفة، مع إيلاء اهتمام خاص لآليات الدمج المتأخر بوصفها خيارًا عمليًا يتيح الجمع بين مصادر متعددة للسمات دون فرض قيود صارمة على طبيعة البيانات أو تزامنها الزمني. كما ناقشت الدراسة إشكاليات جوهرية تتعلق بانحياز الكاميرات، تفاوت جودة البيانات، وحدود قابلية تعميم النماذج عند الانتقال من البيئات المعيارية إلى السيناريوهات الحقيقية.

تستعرض هذه الأطروحة تجارب مقارنة عملية عديدة لتقييم خيارات النماج المستخدمة سواء في استخراج المعطيات السبقية كالصور الظلية واستخراج الوضعيات أو دراسة مقارنة للنماذج اللغوية ضمن سياق المسألة، بهدف مفاضلة استخدام الخيارات المتاحة من حيث الدقة وسرعة الأداء.

وأظهرت النتائج أن الاعتماد على تمثيل أحادي يؤدي غالبًا إلى أداء محدود أو غير مستقر، في حين أن المقاربات الهجينة التي تدمج المظهر، الحركة، والدلالة تحقق توازنًا أفضل بين الدقة والمرونة التشغيلية. كما بيّنت الدراسة أن السمات الدلالية المستخلصة عبر النماذج اللغوية—البصرية تمثل إضافة واعدة، لكنها تتطلب تصميمًا معماريًا داعمًا لتقليل الأخطاء الدلالية وضبط كلفتها الحسابية.

ختامًا، تقترح هذه الأطروحة إطارًا تحليليًا يربط بين طبيعة التمثيلات، استراتيجيات الدمج، وشروط التشغيل الواقعية، وتضع مجموعة من الآفاق المستقبلية والتوصيات البحثية والهندسية، من بينها دمج ذكاء الوكلاء، تحسين بنى التخزين والاسترجاع، وتطوير نماذج أكثر قابلية للنشر في بيئات الحوسبة الحافّية. وتسهم هذه الدراسة في توفير رؤية شمولية تساعد في توجيه الأبحاث المستقبلية نحو أنظمة إعادة تعرّف أكثر موثوقية، تفسيرية، وقابلية للتعميم.

Abstract

This study addresses the problem of person re-identification as a multi-dimensional challenge ,in which appearance-based, temporal, kinematic, and semantic characteristics intersect, and which cannot be adequately captured by a single representation or a single processing pathway. This thesis is a rich and comprehensive reference on the subject of re-identification, covering the most popular and widely used re-identification methods in recent years in a broad and detailed manner. The study aims to provide a critical analysis of person re-identification approaches based on appearance, gait, and pose, in addition to recent methods that leverage vision–language models to extract high-level semantic features, while highlighting the strengths and operational limitations of each category within real-world surveillance contexts.

The adopted methodology is based on an in-depth literature review supported by practical experiments, focusing on evaluating the effectiveness of different fusion strategies, with particular emphasis on late fusion mechanisms as a practical choice that enables the integration of multiple feature sources without imposing strict constraints on data modality or temporal synchronization. The study also discusses fundamental challenges related to camera bias, data quality variability, and the limited generalization capability of models when transitioning from benchmark datasets to real-world scenarios. This thesis reviews several comparative practical experiments to evaluate the modeling options used in extracting prior data such as silhouettes and positions, or in a comparative study of linguistic models within the context of the issue, with the aim of comparing the available options in terms of accuracy and speed of performance.

The results indicate that reliance on a single representation often leads to limited or unstable performance, whereas hybrid approaches that integrate appearance, motion, and semantic information achieve a better balance between accuracy and operational flexibility. Furthermore, the study shows that semantic features extracted via vision–language models constitute a promising addition, but require supportive architectural design to mitigate semantic errors and control computational cost.

In conclusion, this thesis proposes an analytical framework that links representation types, fusion strategies, and real-world operational constraints, and presents a set of future directions and research and engineering recommendations, including the integration of agent-based intelligence, improvements in storage and retrieval architectures, and the development of models that are more suitable for deployment

in edge-computing environments. This study contributes a comprehensive perspective that helps guide future research toward more reliable, interpretable, and generalizable person re-identification systems.

الفهرس

الفصل الأول: مقدمة عامة 1

1.1 مقدمة 1

1.2 أنظمة المراقبة بالكاميرات 1

1.3 أنظمة المراقبة الذكية الموزعة (IDVSS) والمعطيات الكبيرة (Big Data) 4

1.4 أنظمة إعادة المطابقة Re-identification systems 7

1.5 السمات الحيوية (Biometrics) وتحديد الهوية 8

1.5.1 تعريف 8

1.5.2. الركائز السبع للقياسات الحيوية 8

1.5.3. جدول مقارن للخصائص البيومترية 9

1.6. المشكلة العلمية في مشروع البحث 10

1.7 الدافع من البحث 11

1.8 أهمية البحث 11

1.9 مقاربة الحل والسمات المميزة المعتمدة 12

1.9.1. المظهر Appearance 12

1.9.2. بيانات تدفقية 12

1.9.3. سمات دلالية ولونية 13

1.10 المساهمات الأساسية للبحث 13

1.11 بنية الأطروحة 13

الفصل الثاني: الدراسة النظرية 16

.1.2مقدمة 16

.2.2 نماذج استخراج الوضعيات Pose Extraction Models 16

2.2.1. وضعية الجسم كبنية معطيات 16

2.3. نمط المشي Gait 19

2.4. الشبكات العصبونية البيانية (Graph Neural Networks)[2] 21
2.5. شبكات الانتباه البيانية (Graph Attention Networks – GATs)[3] 22
2.6. المحولات [4] Transformers 23
2.7. الشبكات العصبونية البصرية القائمة على بنية المحوّل [5](Vision Transformer) 25
2.8. محوِّل النوافذ المُزاحة [6] (Shifted Window Transformer – Swin) 26
2.9. نماذج اللغة الكبيرة ونماذج الرؤية–اللغة LLMs and VLMs 27
2.10. المعايرة الدقيقة للنماذج اللغوية الكبيرة Fine Tuning LLMs 28
2.11. الحوسبة الحافية Edge Computing 29
2.12. تمثيل COCO-17 لمفاصل جسم الإنسان في تقدير الوضعية البشرية [8] 30
الفصل الثالث: الدراسة المرجعية 32
3.1. مقدمة 32
. 3.2. أنماط مجموعات البيانات المستخدمة في مسألة إعادة المطابقة 32
3.2.1. مجموعات إعادة التعرّف المعتمدة على الصور (Image-Based Appearance Re-ID) 32
3.2.2. مجموعات البيانات الموجّهة للمشية (Gait-Oriented Datasets) 34
3.3. مجموعات البيانات المعيارية المتعلقة بمسألة إعادة المطابقة 35
3.3.1. صورة عامة عن مجموعات البيانات المعيارية ذات الصلة بمسألة إعادة المطابقة 35
3.3.2. مجموعة بيانات 1501-Market[9] 39
.3.3.3 مجموعة بيانات DukeMTMC-reID [10] 40
3.3.4. مجموعة بيانات DukeMTMC-VideoReID [11] 41
3.3.5. مجموعة بيانات (MARS: Motion Analysis and Re-identification Set) [12] 42
3.3.6. مجموعة بيانات (ICFG-PEDES (Identity-Centric and Fine-Grained PEDES [13] 43
3.3.7. مجموعة بيانات CUHK03 [14] 44
3.3.8. مجموعات بيانات المشية لإعادة التعرف 45
3.4. المشية والخصائص البيومترية المرتبطة بها 50
3.4.1. مقدمة 50
3.4.2. التغيّر في وضعية الجسم (Pose) خلال المشي كخاصية مميّزة 51
3.4.3. التغير في شكل الجسم (الصورة الظلية) كخاصية مميزة في مسألة إعادة التعرف 58
3.4.4. النماذج الهجينة في التعرّف على المشية 67
3.5. خصائص الهيئة والمظهر (Appearance) في مسألة إعادة التعرف 70
3.6 السمات الدلالية في مسألة إعادة التعرف 81

3.6.1. نموذج CLIP-REID [28] 81
3.6.2. اطار عمل LVLM_ReID [29] 83
الفصل الرابع: الإطار العملي 89
4.1 مقدمة 89
4.2 استخراج المعطيات المناسبة كمدخلات لنماذج إعادة التعرف المعتمدة على المشية 89
4.2.1. استخراج الوضعيات Pose Extraction 90
4.2.2. استخراج الصور الظلية 94
4.3. إشكالية توافر مجموعات البيانات الشاملة لتقييم نماذج إعادة التعرّف المعتمدة على المشية 102
4.4. نموذج القناة الحركية (Kinematic Stream) 105
4.4.1. تمثيل البيانات ومدخلات النموذج 106
4.4.2. **البنية الزمنية وبناء المتتبعات (Skeleton Tracklets)** 107
4.4.3. تهيئة المدخلات لنمذجة بنيوية–زمنية 107
4.4.4. المعمارية الأساسية لنموذج KinematicGNN 108
4.5. النموذج الهجين المعتمد على المظهر Hybrid Appearance Transformer (HAT-ReID) 118
4.6. النموذج الوصفي الدلالي 120
4.6.1. التوصيف النصي لصور المشاة كخاصية دلالية 120
4.6.2. التحول من توصيف لغوي صحيح إلى تمثيل دلالي تمييزي 128
4.7. دمج الخصائص Fusion Embeddings 132
4.7.1. دمج النماذج في أنظمة إعادة التعرّف 132
الفصل الخامس: الخاتمة والآفاق المستقبلية 138
5.1. الخاتمة 138
5.2. الآفاق المستقبلية (Research Directions) 138
5.3. التوصيات الهندسية (Engineering Recommendations) 139
المراجع 141

جدول الأشكال

الشكل 1 إعادة التعرف لمرور شخص أمام عدة كاميرات في اوقات مختلفة 7
الشكل 2 أطوار عملية المشي عند الإنسان 20
الشكل 3 شبكات الانتباه البيانية Graph Attention Networks 22
الشكل 4 معمارية المحولات Transformers 24
الشكل 5 بنية المحول البصري Vision Transformer ViT 25
الشكل 6 اعادة ترتيب النوافذ في خوارزمية swin 26
الشكل 7 البنية الهرمية في swin 26

الشكل 8 نموذج لغوي بصري.....27
الشكل 9 مبدأ عمل خوارزمية LoRA.....28
الشكل 10 تمثيل Coco-17 لوضعية جسم الإنسان.....31
الشكل 11 مجموعة بيانات CASIA-B الزوايا وطريقة التقاط الصور.....46
الشكل 12مجموعة بيانات OU-MVLP الزوايا وطريقة التقاط الصور.....48
الشكل 13 المعطيات الظلية والوضعيات في الفضائين ثنائي الابعاد وثلاثس الابعاد المشتقى من صور محموعة البيانات GREW.....49
الشكل 14 تمثيل بيان وضعية جسم الانسان في مصفوفة تجاور.....51
الشكل 15 بنية نموذج posegait.....53
الشكل 16بنية نموذج gaitgraph.....54
الشكل 17 بنية ومبدأ عمل GaitPart.....60
الشكل 18 الرقع عديمة القيمة في الصور الظلية.....65
الشكل 19 الهيكل والوضعية مبنية باستخدام الخرائط الحرارية وفقاً لمقاربة SkeletonGait.....68
الشكل 20 مبدأ عمل Omnis-Scale Network.....76
الشكل 21 بنية TE-TransReID.....80
الشكل 22 بنية وطريقة عمل CLIP-ReID.....82
الشكل 23 بنية وآالية عمل إطار العمل Lvlm-ReID.....84
الشكل 24 إحدى العينات ونتائج استخراج الصورة الطلية منها بالطرائق المختبرة - تظهر بوضوح عدم جودة النتائج.....100
الشكل 25 إحدى العينات والصور الظلية المستخرجة منها.....103
الشكل 26 Casia - B و طريقة التقاط صورها والظلال المستخرجة من عدة زوايا.....104
الشكل 27 عينة من البيانات تشرح الصيغة الجدولية لنسخة casia-b بظهر فيها كل نقطة مع إحداثياتها.....106
الشكل 28 التمثيبلات البيانية المستخرجة من العقد واحداثياتها فوق صيغة coco-17 تنتمي الصور إلى عنصر المرور ذاته.....106
الشكل 29 معمارية KinematicGNN.....109
الشكل 30 يوضح الشكل كيف بفصل Arcface الفضاء الهوياتي بشكل أكثر وضوحاً من softmax.....113
الشكل 31 معمارية النموذج الهجين المعتمد على المظهر.....119
الشكل 32 النموذج اللغوي Smolvlm2 + PSTG.....130
الشكل33 شعاع دمج مركب من عدة أشعة ناتجة عن نماذج مختلفة وبأبعاد مختلفة.....133

جدول الجداول

الجدول1 الجواسب أحادية اللوحة.....6
الجدول2 جدول مقارن للخصائص البيومترية.....10
الجدول 3 مقارنة بين hrenet و yolo-pose.....19
الجدول4 جدول المقارن بين ـأنماط مجموعات البيانات.....35
الجدول 5 أجيال النماذج المعتمدة على تغير الوضعيات ومقارباتها.....56
الجدول6 أجيال النماذج المعتمدة على تغير الوضعيات البتية وتمثيل المعطيات.....57
الجدول 7 أجيال النماذج المعتمدة على تغير الوضعيات مقارنة جودة النتائج.....57
الجدول 8 أجيال النماذج المعتمدة على تغير الوضعيات والقيمة المضافة والتحديات.....57
الجدول 9 مقارنة النماذج المعتمدة على الصور الظلية على مجموعة بيانات casia-b.....66
الجدول 10 الجدول 9 مقارنة النماذج المعتمدة على الصور الظلية على مجموعات بيانات المشية.....67
الجدول 11 مقارنة SkeletonGait مع نماذج الصور الظلية على محموعة البيانات Casia-b.....70
الجدول 12 مقارنة SKELETONGAIT مع نماذج الصور الظلية على محموعة البياناتGrew.....70
الجدول 13 جدول مقارنة ين CLIP-ReID و LVLM-ReID.....87
الجدول 14 ين CLIP-ReID و LVLM-ReID من ناحية مجموعات البيانات.....87
الجدول15 ين CLIP-ReID و LVLM-ReID على صعيد الجودة.....87
الجدول 16 مقارنة أداء طرائق استخراج الصور الظلية على صعيد زمن التنفيذ.....96
الجدول17 مقارنة أداء طرائق استخراج الصور الظلية على صعيد مقاييس الجودة.....99
الجدول 18 تطور معمارية kienmaticGNN والنمتائج في كل مرحلة.....116
الجدول19 مقارنة نتائج تحارب النماذج المعتمكدة على المظهر ومن ضمنها نموذجنا HAT.....119
الجدول 20 مقارنة النموذجين البصريين QWEN و LLAVAمع نمطي التلقين البسيط والمعقد.....123

الجدول21 احتبار أداء Smolvlm2-500M.....126
الجدول22 جدول نتائج تجربة خيارات الإطار الممثل للسلسلة مع متوسطها الحسابي.....134
الجدول23 جدول نتائج تجارب تطبيق الدمج بين نموذجين.....136

المصطلح الإنجليزي	المقابل العربي
Person Re-Identification (Re-ID)	إعادة التعرّف على الأشخاص
Re-Matching	إعادة المطابقة
occulasions	حجب
Biometrics	الخصائص المميّزة البيومترية
Appearance Features	الخصائص المميّزة المظهرية
Kinematic Features	الخصائص المميّزة الحركية
Semantic Features	الخصائص المميّزة الدلالية
Gait	المشية
Gait Cycle	دورة المشي
Pose	الوضعية
Skeleton	الهيكل العظمي
Joint	المفصل
Silhouette	الصورة الظلية
Tracklet	عنصر مرور

إطار (زمني)	Frame
تسلسل زمني	Sequence
تمثيل بياني	Graph Representation
شبكة التفاف بيانية	Graph Convolutional Network (GCN)
شبكة التفاف عصبونية	Convolutional Neural Network (CNN)
شبكة عصبونية تكرارية	Recurrent Neural Network (RNN)
النمذجة الزمنية	Temporal Modeling
آلية الانتباه	Attention Mechanism
المحوّل	Transformer
المحوّل البصري	Vision Transformer (ViT)
المحوّل ذو النوافذ المنزاحة	Swin Transformer
فضاء الأشعة	Embedding Space
شعاع تمثيلي	Embedding Vector
تابع الخسارة الثلاثية	Triplet Loss Function
تابع أرك فيس الزاوي	ArcFace Loss Function
عنق التطبيع الدفعي	Batch Normalization Neck (BNNeck)
نموذج لغوي–بصري كبير	Large Vision-Language Model (LVLM)

إطار إعادة التعرّف اللغوي–البصري	LVLM-ReID Framework
رمز دلالي	Semantic Token
الدمج المبكر	Early Fusion
الدمج المتأخر	Late Fusion
دمج الأشعة	Embeddings Fusion
تحيز الكاميرا	Camera Bias
قابلية التعميم	Generalization
دقة الرتبة الأولى	Rank-1 Accuracy
دقة الرتب المتعددة	Rank-5 Accuracy
متوسط الدقة المتوسطة	mean Average Precision (mAP)
بيئة معطيات كبيرة	Big Data Environment
حوسبة حافية	Edge Computing
نظام مراقبة موزع	Distributed Surveillance System
نظام التلفزة ذات الدارة المغلقة	Closed-Circuit Television (CCTV)
مجموعة بيانات معيارية	Benchmark Dataset
مجموعة بيانات ماركت-1501	Market-1501
مجموعة بيانات ديوك MTMC لإعادة التعرّف	DukeMTMC-ReID

مجموعة بيانات مارس	MARS Dataset
مجموعة بيانات كاسيا-بي	CASIA-B
مجموعة بيانات أو-إم في إل بي	OU-MVLP
مجموعة بيانات غرو	GREW Dataset
مجموعة بيانات ICFG-PEDES	ICFG-PEDES

الفصل الأول: مقدمة عامة

1.1 مقدمة

يحتل "الأمان والاستقرار" (Safety Needs) المرتبة الثانية في هرم ماسلو للحاجات الإنسانية، مباشرة بعد إشباع الحاجات الفيزيولوجية الأساسية (كالطعام والشراب). تمثل هذه الحاجة دافعاً إنسانياً أصيلاً للبحث عن بيئة خالية من الفوضى والتهديدات، وتوفير الأمان الجسدي، والوظيفي، والصحي، وأمن الممتلكات. ولتحصيل هذا الأمان، اتبع الإنسان سُبلاً متعددة عبر تاريخه، بدءاً من بناء الملاجئ والأسوار، ومروراً بتأسيس أنظمة القانون والشرطة لفرض النظام الجماعي. وتُعتبر أنظمة المراقبة المعتمدة على الكاميرات التجسيد التكنولوجي الحديث لهذا السعي الفطري؛ فهي تعمل كـ "رادع" (Deterrent) يمنع التهديدات المحتملة، وتمنح الفرد شعوراً بالسيطرة والاطمئنان (راحة البال) من خلال المراقبة اللحظية لمحيطه وممتلكاته، وتوفر دليلاً للرد على الانتهاكات، مشبعةً بذلك إحدى أعمق الحاجات الإنسانية الأساسية التي لا غنى عنها للانتقال إلى المستويات الأعلى من تحقيق الذات، ومنه كان من البديهي أن يواكب تطور التقانة في مختلف مناحي الحياة تطوراً في هذا الجانب الحيوي والمهم من حياة البشر، ويعد ظهور أنظمة المراقبة من أهم التطبيقات المباشرة للتقانة على هذا الصعيد ومن أكثر الأنظمة استخداماً وانتشاراً اليوم على مختلف المستويات بدءاً من من تطبيقها على مستوى متجر صغير ووصولاً إلى مستوى المدن والأقاليم.

1.2 أنظمة المراقبة بالكاميرات

لطالما مثّلت أنظمة المراقبة المعتمدة على الكاميرات (CCTV) خط الدفاع الأول في تأمين الممتلكات والمساحات العامة. لكن دورها تطور بشكل جذري، فبعد أن كانت مجرد أدوات تسجيل سلبية، تحولت اليوم إلى أنظمة تحليلية معقدة تُنتج كميات هائلة من المعطيات، ولا يمكن استغلالها بفعالية إلا بدمجها مع تقنيات الذكاء الصنعي.

الأهمية المتنامية والاحتياجات الملباة

في بدايات ظهورها، كان الهدف الأساسي لأنظمة المراقبة بسيطاً: الردع والتوثيق. مجرد وجود كاميرا كان كافياً لردع بعض المجرمين، وفي حال وقوع حادث، كانت التسجيلات تُستخدم كدليل بعد وقوع الحدث (Post-event analysis). كانت الاحتياجات التي تلبيها تقتصر على:

1. الأمن الأساسي: مراقبة المداخل والمخارج والمناطق الحساسة.
2. التوثيق الجنائي: توفير دليل مرئي للشرطة وجهات التحقيق.
3. المراقبة البشرية المباشرة: وجود حارس أمن يراقب عدة شاشات في غرفة تحكم.

أما اليوم، فقد اتسعت دائرة الاحتياجات بشكل هائل لتشمل الاستباقية والكفاءة التشغيلية. لم تعد الأنظمة تُستخدم للأمن فقط، بل أصبحت تلبي احتياجات تحليلية وتشغيلية معقدة في مختلف القطاعات.

تساعدنا أنظمة المراقبة في مواجهة طيف واسع من التحديات، تتجاوز مجرد رصد السرقات:

تحديات أمنية

- الاستجابة الفورية: الانتقال من "ماذا حدث؟" إلى "ماذا يحدث الآن؟".
- التحليل الاستباقي: رصد السلوكيات المشبوهة (مثل التسكع في منطقة محظورة، أو ترك حقيبة غريبة).
- إدارة الحشود: رصد الكثافات البشرية الخطرة والتنبيه قبل حدوث تدافع.
- التعرف على الهويات: التحقق من قوائم المطلوبين أو إدارة التحكم في الدخول (Access Control).

تحديات غير أمنية (تشغيلية وتحليلية)

- تحسين تجربة العملاء (في التجزئة): تحليل مسارات العملاء (Heatmaps) لمعرفة أكثر الأقسام زيارة، وتحديد أوقات الذروة لتقليل طوابير الانتظار.
- إدارة المرور: رصد الاختناقات المرورية آلياً، وقراءة لوحات المركبات (ANPR) لإدارة المواقف أو المخالفات.
- السلامة المهنية (في المصانع): التأكد من ارتداء العمال لمعدات السلامة (الخوذ والقبعات)، ورصد الحوادث (مثل سقوط شخص) وإرسال تنبيه فوري.
- مراقبة الجودة: فحص المنتجات على خطوط الإنتاج بصرياً وبسرعة تفوق القدرة البشرية.

جدوى الأنظمة التقليدية وصعوبة الاستفادة منها

إن الاعتماد على أنظمة المراقبة بصيغتها التقليدية (أي مجرد تسجيل الفيديو على أجهزة DVR/NVR) دون أنظمة استرجاع ذكية، يجعلها منخفضة الجدوى بشكل كبير في البيئات التي تحتوي على أكثر من بضع كاميرات.

الصعوبة تكمن في "البحث عن إبرة في كومة قش".

لنفترض على سبيل المثال أننا نريد استرجاع معلومة تخص دخول شخص يرتدي معطفاً أحمر إلى مبنى ما خلال الـ 24 ساعة الماضية، ولدينا 50 كاميرا ضمنه، في النظام التقليدي، سيتعين على موظف أمن مراجعة 24 ساعة من التسجيلات لـ 50 كاميرا (أي 1200 ساعة من الفيديو) يدوياً. هذا مستحيل عملياً، ومكلف، ويأتي دائماً متأخراً جداً.

الأنظمة التقليدية فعالة فقط كـ "صندوق أسود" يُرجع إليه بعد تحديد وقت ومكان الحادث بدقة، لكنها تفشل تماماً في الاستكشاف أو الاستجابة اللحظية.

عندما تتحول الكاميرات من مجرد "تسجيل" إلى "تحليل"، فإنها تنتج بيانات غنية يمكن تحويلها إلى تقارير ومعارف قيمة:

- المعطيات الأولية (Raw Data): فيديو، صور، إحداثيات الأجسام المتحركة.
- المعطيات الوصفية (Metadata):
  - الإحصاء (Counting): عدد الأشخاص، عدد المركبات.
  - التصنيف (Classification): تمييز (شخص، سيارة، دراجة، حيوان).
  - السمات (Attributes): لون الملابس، نوع المركبة (شاحنة، سيارة ركاب)، اتجاه الحركة.
  - الأحداث (Events): عبور خط افتراضي، دخول منطقة محظورة، توقف مركبة.
- التقارير والمعارف (Reports & Insights):
  - تقارير أمنية: ملخص بالتنبيهات (مثل محاولات تسلق الأسوار)، أوقات حدوثها، ومواقعها.
  - تقارير تشغيلية: خرائط حرارية (Heatmaps) توضح أكثر المسارات ازدحاماً.
  - تحليلات سلوكية: متوسط وقت بقاء العميل في منطقة معينة.
  - بحث ذكي: القدرة على البحث بـ "أرني كل السيارات الحمراء التي مرت بهذا الشارع بين 3-4 عصراً".

حتمية تضمين الذكاء الصنعي في أنظمة المراقبة

هنا يكمن المفصل الحقيقي. الذكاء الصنعي (وتحديداً خوارزميات تعلم الآلة والتعلم العميق في مجال الرؤية الحاسوبية – Computer Vision) هو المحرك الذي يحول الفيديو من بيانات ميتة إلى معلومات حية.

تكمن أهمية تضمين الذكاء الصنعي في:

1. الأتمتة (Automation): بدلاً من أن يراقب الإنسان 100 شاشة، يقوم الذكاء الصنعي بمراقبتها جميعاً في نفس اللحظة، ولا ينبه الإنسان إلا عند وقوع حدث مهم.
2. السرعة (Speed): تحليل آلاف الساعات من الفيديو المسجل في دقائق معدودة للعثور على لقطة معينة (Smart Search).
3. الدقة (Accuracy): قدرة الخوارزميات على رصد تفاصيل دقيقة (مثل قراءة لوحة سيارة مسرعة) تفوق قدرة العين البشرية.
4. الاستباقية (Proactivity): التعرف على الأنماط الخطرة (مثل تجمع مريب لأشخاص) والتنبيه *قبل* وقوع الجريمة.

1.3 أنظمة المراقبة الذكية الموزعة (IDVSS) والمعطيات الكبيرة (Big Data)

تُعرّف أنظمة المراقبة الذكية الموزعة (Distributed Intelligent Surveillance Systems) بأنها بنية شبكية متقدمة للمراقبة، لا تعتمد على خادم مركزي واحد لمعالجة وتحليل جميع بيانات الفيديو

إنّ أنظمة المراقبة الذكية الموزعة هي أنظمة "بيانات ضخمة" (Big Data) حيث أن

المعطيات الناتجة عن أنظمة المراقبة الحديثة هي مثال كلاسيكي على "المعطيات الضخمة" (Big Data)، وهي تحقق بوضوح المعايير الثلاثة الرئيسية (The 3Vs):

1. الحجم (Volume)

حجم المعطيات هائل. كاميرا واحدة بجودة HD أو K4 تسجل 7/24 يمكن أن تنتج عشرات أو مئات الجيجابايت يومياً. في نظام مراقبة لمدينة ذكية أو مركز تجاري كبير (يضم 500 أو 1000 كاميرا)، نتحدث عن "بيتا بايتات" (Petabytes) من المعطيات التي تحتاج إلى تخزين ومعالجة.

2. السرعة (Velocity)

وهي السمة الأكثر تحدياً. المعطيات ليست ضخمة فقط، بل هي متدفقة (Streaming) بسرعة لحظية.

- مثال: لنفترض أن لدينا كاميرا تلتقط 30 إطاراً في الثانية (FPS 30). هذا يعني أن النظام يجب أن يستقبل ويعالج 30 صورة *كل ثانية من هذه الكاميرا الواحدة*.
- إذا كان النظام يضم 100 كاميرا، فهو يستقبل 3000 إطار (صورة) في الثانية الواحدة.
- التحدي (القيد الزمني) هو أن تحليل هذه الأطر يجب أن يتم في الزمن الحقيقي (Real-time). لا فائدة من التنبيه عن حادث سرقة بعد 10 دقائق من وقوعه. يجب أن يتم التحليل (مثل رصد حركة، التعرف على وجه) في أجزاء من الثانية (milliseconds) لتكون الاستجابة فعالة.

3. التنوع (Variety)

المعطيات ليست مجرد "فيديو" موحد. هي بيانات غير مهيكلة (Unstructured) ومتنوعة للغاية:

- تنوع المصادر: لقطات من كاميرات نهارية، كاميرات ليلية (تحت حمراء)، كاميرات حرارية (Thermal)، كاميرات بزوايا مختلفة (عين السمكة، بانورامية).
- تنوع المحتوى: الفيديو قد يحتوي على وجوه بشرية، لوحات مركبات، نصوص (على لافتات)، أصوات (إذا كانت الكاميرا تحوي ميكروفوناً)، بالإضافة إلى بيانات وصفية (Metadata) مثل الموقع الجغرافي (GPS) والوقت.

هذا المزيج من الحجم الهائل، والتدفق اللحظي فائق السرعة، والتنوع الكبير في المعطيات، يجعل من المستحيل إدارتها أو تحليلها بالطرق التقليدية، ويفرض حتمية استخدام بنى تحتية للبيانات الضخمة وخوارزميات الذكاء الصنعي لاستخلاص القيمة منها.

إن تعددية مصادر المعطيات وضخامتها والمقاربة المعتمدة على الحوسبة الموزعة يجعل من طرح الحوسبة الحافية في هذا السياق أمراً منطقياً إذ يكاد يكون هذا النوع من الأنظمة الحاسوبية الحالة المثلى لتطبيقها.

الحوسبة الحافية (Edge Computing) ودور الحواسب الصغرية

هي نموذج حوسبة موزع، يقوم بنقل عمليات المعالجة (مثل خوارزميات الذكاء الصنعي) من الخوادم المركزية البعيدة (مثل السحابة) إلى أقرب نقطة ممكنة من مصدر إنتاج المعطيات.

"الحافة" (The Edge) هي الجهاز المادي نفسه، والذي أصبح ممكناً بفضل الحواسب الصغرية (Microcomputers).

فبدلاً من الاعتماد على خوادم ضخمة، تُدمج وحدات معالجة قوية ومصغرة داخل الكاميرا الذكية أو جهاز التسجيل المحلي (NVR). تُستخدم هذه الحواسيب الصغرية على نطاق واسع في تطبيقات الحوسبة الطرفية وتأتي بأشكال متنوعة، منها:

- **الحواسيب أحادية اللوحة (SBCs)**: وهي لوحات حاسوبية متكاملة، أبرزها وأكثرها شيوعاً

Raspberry Pi	الأكثر شيوعاً، تعتمد على وحدات معالجة تقليدية في معظم النماذج، والحديثة منها تحتوي على معالجات رسومية متوسطة الأداء غالباً ما يترافق استخدمها مع مسرعات أداء مثل وحدات coral
Khadas Vim	أكثر متانة من Raspberry Pi وتعتمد في نماذجها على تضمين وحدات معالجة ذات معماريات متوافقة مع الشبكات العصبونية NPU فهي في أساس تصميمها موجهة لاستثمارها في الأنظمة الذكية على عكس Raspberry Pi ذات التوجه العمومي متعدد الأغراض
Radaxa Rock	هذه السلسلة من الشركة الرائدة اليوم في عتاديات الرسوميات والذكاء الصنعي تعتمد على معالجات رسوميات قوية GPU وتوافقية مع أطر العمل الشهيرة مثل Cuda
Nvidia Jetson	هذه السلسلة من الشركة الرائدة اليوم في عتاديات الرسوميات والذكاء الصنعي تعتمد على معالجات رسوميات قوية GPU وتوافقية مع أطر العمل الشهيرة مثل Cuda

الجدول1 الجواسب أحادية اللوحة

- **المتحكمات الدقيقة مثل (Arduino MCUs أو 32ESP)**: وهي أنظمة أبسط تُستخدم لمهام التحكم وقراءة الحساسات الأساسية.

- **مسرّعات الذكاء الصنعي(AI Accelerators) :** ك وحدات (Google Coral Edge TPU) التي تُدمج مع الحواسب الصغرية لتسريع التحليلات الذكية وتعتمد وحدات معالجة من نوع خاص بحسابات التنسورات Tensors وهو أحد أكثر

أشكال تمثيل المعطيات والخصائص المستخرجة في أنظمة الذكاء الصنعي وتعلم الآلة وخاصة تلك المعتمدة على الرؤية الحاسوبية.

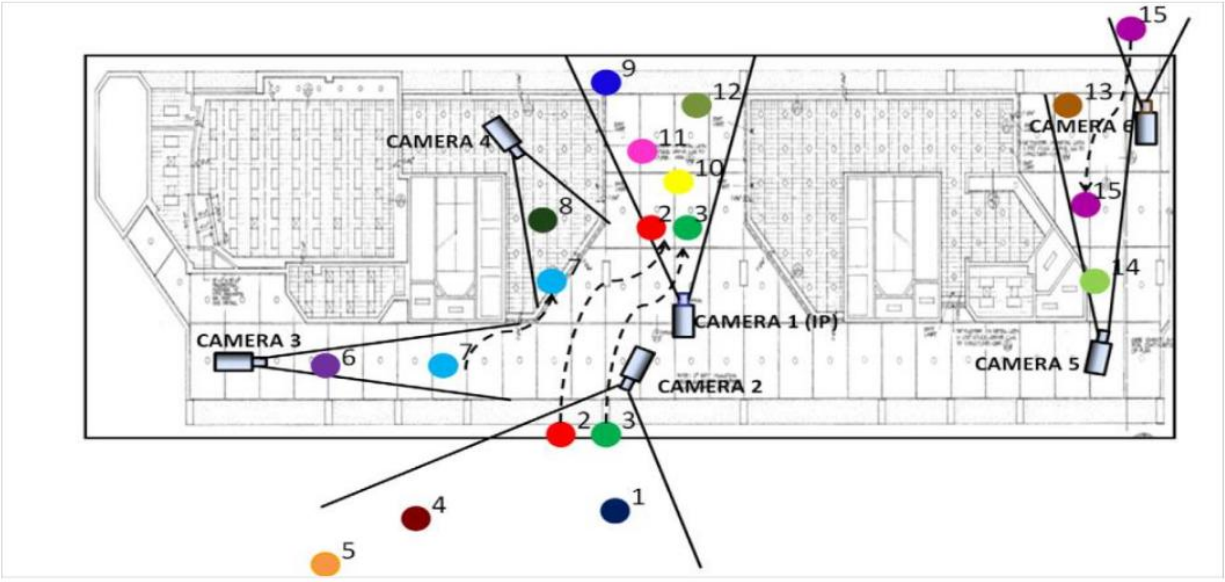
بدلاً من إرسال تدفق الفيديو الخام بالكامل عبر الشبكة للتحليل المركزي، تقوم هذه الحواسيب الطرفية بتحليل الفيديو محليًا، وترسل فقط النتائج النهائية (مثل التنبيهات أو المعطيات الوصفية). يضمن هذا النهج استجابة سريعة للغاية (Low Latency) ويقلل بشكل كبير الضغط على شبكة الاتصالات.

1.4 أنظمة إعادة المطابقة Re-identification systems

يمكن تعريف عملية إعادة المطابقة re-identification كالتالي:

هي عملية إثبات أن كائناً ما هو "نفسه" الذي شوهد في وقت أو مكان آخر، *حتى لو كانت هويته (اسمه) مجهولة تماماً*.

مثال: نظام "إعادة تحديد الهوية" (Re-Identification). يرى النظام شخصاً في الكاميرا (1). بعد 10 دقائق، يرى شخصاً الكاميرا (5). يقوم النظام بعملية "مطابقة" ليقول "هذا الشخص هو نفس الشخص الذي كان هناك"، دون أن يعرف بالضرورة أن اسمه "أحمد". اي أنها تجيب على سؤال: "هل هذا نفس الشخص الذي رأيته من قبل؟ وليس سؤال: "هل أعرف **من** هذا؟" وهو سؤال إعادة التعرف وليس إعادة المطابقة.



الشكل 1 إعادة التعرف لمرور شخص أمام عدة كاميرات في اوقات مختلفة

يوضح الشكل نظام مراقبة موزع يستخدم ست كاميرات وكيفية التعرف على الأشخاص وتتبع مسارات حركتهم من خلال الربط والمطابقة بين المعطيات الخاصة بكل مرور أمام إحدى الكاميرات، أما فيما يخص ماهية هذه المعطيات والخصائص فهي عديدة منها ما يندرج في إطار السمات الحيوية أو البيومترية، ومنها ما مختلف عنها.

1.5 السمات الحيوية (Biometrics) وتحديد الهوية

1.5.1 تعريف

يمكننا تعريفها على أنها هي الخصائص البيولوجية والسلوكية القابلة للقياس والمميزة لشخص ما عن آخر.

1.5.2. الركائز السبع للقياسات الحيوية

لكي تعتبر أي سمة بيولوجية أو سلوكية طريقة قياس حيوي قابلة للتطبيق، يجب أن تمتلك مجموعة من الخصائص الأساسية. هذه الخصائص، التي يشار إليها غالبًا بالركائز السبع للقياسات الحيوية، تشكل إطاراً لتقييم مدى ملاءمة طريقة ما لتطبيق معين وهي:

1. الشمولية (Universality): يجب أن يمتلك كل فرد ضمن السكان المستهدفين الذين يستخدمون النظام هذه السمة.
2. التفرد (Uniqueness/Distinctiveness): يجب أن تكون السمة مختلفة بما فيه الكفاية عبر الأفراد في السكان للسماح بالتمييز القاطع. على سبيل المثال، تظهر الأنماط المعقدة للقزحية تفردًا عاليًا للغاية، حتى بين التوائم المتطابقة.
3. الثبات (Permanence): يجب أن تظل سمة القياس الحيوي ثابتة بما فيه الكفاية على مدى حياة الشخص، أو على الأقل لفترة الاستخدام ضمن التطبيق.
4. قابلية الجمع (Collectability/Measurability): يجب أن يكون من الممكن الحصول على السمة ورقمنتها باستخدام تقنية استشعار مناسبة.
5. الأداء (Performance): يجب أن تفي قدرة النظام على التعرف على السمة بالمعايير المطلوبة للدقة والسرعة وكفاءة الموارد للتطبيق المقصود.
6. القبول (Acceptability): يجب أن يكون الأفراد المستهدفين على استعداد لتقديم سمتهم الحيوية للنظام، وهنا يتم التمييز بين أفراد متعاونين وأفراد غير متعاونين أو خصائص يتم جمعها بشكل سلبي أو بشكل تفاعلي بين الفرد والمنظومة.
7. التحايل (Circumvention): يجب أن تكون السمة صعبة التقليد أو الخداع باستخدام أدوات مصطنعة.

يمكن تقسيم السمات الحيوية إلى قسمين:

- سمات جسدية أو فيزيولوجية: وهي عوامل فطرية تتحدد بشكل أساسي بمزيج العوامل الوراثية الخاصة بالفرد وقد تطرأ عليها تغيرات بسيطة مع التقدم بالعمر.
- سمات سلوكية: على النقيض من الطبيعة الثابتة للسمات الجسدية، تقيس القياسات الحيوية السلوكية الأنماط الديناميكية الفريدة الكامنة في تصرفات الفرد وتفاعلاته. هذه ليست خصائص فطرية ولكنها عادات وميول يتم تعلمها وتطويرها بمرور الوقت من خلال التفاعل المتكرر مع العالم ومع التكنولوجيا. غالبًا ما تكون هذه الأنماط دقيقة ويتم تنفيذها دون وعي، مما يجعل تقليدها صعبًا.

1.5.3. جدول مقارن للخصائص البيومترية

فيما يلي جدول يعدد السمات الحيوية الشهيرة بنوعيها يقارن فيما بينهما على صعيد أهم الركائز.

الطريقة	النوع	التفرد	الثبات	قابلية الجمع	الأداء	القبول	مقاومة التحايل
بصمة الإصبع	جسدي	مرتفع	مرتفع	مرتفع	مرتفع	متوسط–مرتفع	متوسط
الوجه	جسدي	متوسط–مرتفع	متوسط	مرتفع	متوسط	مرتفع	منخفض
القزحية	جسدي	مرتفع جدًا	مرتفع	متوسط	مرتفع جدًا	منخفض–متوسط	مرتفع
نمط الأوردة	جسدي	مرتفع	مرتفع	متوسط	مرتفع	متوسط	مرتفع
الصوت	هجين	منخفض–متوسط	منخفض	مرتفع	منخفض	مرتفع	منخفض
التوقيع	سلوكي	منخفض	منخفض	مرتفع	منخفض	مرتفع	منخفض

ضغطات المفاتيح	سلوكي	منخفض–متوسط	منخفض–متوسط	مرتفع	متوسط	مرتفع	متوسط–مرتفع
المشية	سلوكي	منخفض–متوسط	منخفض	متوسط	منخفض	مرتفع	متوسط
الحمض النووي ((DNA	جسدي	مرتفع جدًا	مرتفع جدًا	منخفض	مرتفع جدًا	منخفض	مرتفع جدًا

الجدول2 جدول مقارن للخصائص البيومترية

1.6. المشكلة العلمية في مشروع البحث

إن مسألة إعادة المطابقة أو إعادة التعرف بشكل فعال من ناحية الدقة وكذلك من ناحية الأداء هي مسألة من الصعوبة بمكان وذلك بسبب عوامل عديدة:

- تعتمد إعادة المطابقة على بيانات يتم جمعها بشكل سلبي، أي أن الأشخاص الذين يتحركون بشكل طبيعي ضمن البيئة التي يعمل ضمنها نظام المراقبة ليسوا حريصين أو معنيين إن لم يكونوا ممانعين في الأساس لتزويد النظام ببيانات ذات جودة كافية.
- تختلف المعطيات التي يمكن جمعها والاستفادة منها اختلافاً كبيراً بين التقاط وآخر للشخص ذاته نذكر من أسباب هذا الاختلاف:
  - زاوية الكاميرا تختلف بحسب موقعها وطبيعة المكان الواقع ضمن قطاع رؤيتها
  - يختلف بعد الشخص عن الكاميرا بين مرور لآخر وكاميرا لأخرى.
  - تختلف ظروف الإضاءة بين كاميرا وأخرى وأحيانا للكاميرا ذاتها بين وقت وآخر.
  - إن كون الأشخاص غير متعاونين مع النظام يؤدي إلى حتمية التعامل مع بيانات ذات إشكاليات مختلفة، كأن يكون جزءاً من صورة الفرد محجوباً بأفراد آخرين قريبين منه أو من عوائق مثل مقعد ما أو شجرة أو سيارة مركونة (occulasions).
  - غالباً ما تكون دقة الصور الملتقطة للأفراد منخفضة بالمقارنة مع أنظمة أخرى تعتمد السمات الحيوية.

- إن السمات الحيوية المعتمدة في هذا النوع من الأنظمة ذات ثبات وتفرد متوسط أو منخفض خاصةً السلوكية منها. الأمر الذي يجعل الاعتماد على سمة واحدة فقط غير مجد في أغلب الحالات حيث أن طبيعة أنظمة المراقبة التي تطبق على بيئات مختلفة تجعل من أية سمة معتمدة ذات جودة مختلفة باختلاف الحالات، للتوضيح نذكر الأمثلة التالية:

  - كاميرا موضوعة في موقع تتعاقب عليه إضاءات مختلفة اختلافات كبيرة، كأن تكون خارج مبنى وبالتالي فالسمات المفيدة لها نهاراً لن تكون بالفائدة ذاتها ليلاً.
  - نظام مراقبة ضمن منشأة ما تفرض زياً موحداً على أفرادها يجعل من محاولة التمييز من خلال الملابس واختلافها غير ذا جدوى.
  - كاميرا موضوعة في بداية ممر تلتقط حركة الأشخاص بحيث يكونوا قادمين باتجاهها أو مبتعدين عنها الأمر الذي يجعل من محاولة بناء نمط يخص المشية مثلا غير مفيد إذا لا تغييرات حقيقية ملتقطة في هذا الخصوص بسبب زاوية الكاميرا ذاتها.

تتناول معظم الأبحاث المنشورة مسألة إعادة المطابقة من وجهة نظر سمة واحدة فقط وتحاول اقتراح نماذج متعلقة بها. وهو كما أسلفنا قاصر من الناحية العملية، على الأقل في بعض الحالات، أي لا يمكن الاعتماد بشكل كلي على سمة واحدة فقط من الناحية العملية عند محاولة تطوير نظام إعادة مطابقة.

1.7 الدافع من البحث

تطوير نموذج لإعادة المطابقة يعتمد على دمج عدة خصائص بيومترية مع مراعاة طبيعة النظام المستهدف كنظام يتعامل مع بيانات كبيرة.

1.8 أهمية البحث

تكمن أهمية هذا البحث في تقديم مقاربة شمولية لمسألة إعادة التعرّف على الأشخاص، من خلال سبر الخصائص المميزة المرتبطة بها، سواء تلك المعتمدة على المظهر، الحركة، المشية، أو السمات الدلالية، وتحليل كيفية الاستفادة من كل منها على نحو مستقل، ثم دراستها ضمن إطار تكاملي يأخذ بعين الاعتبار الأبعاد الزمانية والمكانية للمشهد.

وتنبع القيمة العلمية للبحث من محاولته تجاوز المقاربات الأحادية الشائعة، عبر استكشاف استراتيجيات دمج مرنة تمكّن من تحقيق توازن فعّال بين القدرة التمييزية للنموذج والكلفة الحسابية المفروضة. أما على الصعيد التطبيقي، فيهدف البحث إلى بناء

نموذج إعادة مطابقة ذي كفاءة عالية وقابلية تشغيل واقعية، يمكن دمجه ضمن أنظمة مراقبة ذكية موزعة، مع مراعاة القيود التي تفرضها طبيعة البيانات واسعة النطاق، وعدم تجانس المصادر، ومتطلبات الاستجابة الزمنية.

1.9 مقاربة الحل والسمات المميزة المعتمدة

اتبعنا مقاربة متسقة مع ما هو منشور في الأدبيات ومتسقة مع الحس السليم والحدس والبداهة، فكان أن اعتمدنا على السمات التالية:

1.9.1. المظهر Appearance

إن المعطيات التي تؤمنها أنظمة المراقبة هي بيانات صورية في الأساس، وإذا كان بإمكان أي مشغل لنظام مراقبة مقارنة صورة ملتقطة لشخص ما من إحدى الكاميرات فإنه من البديهي أنه بالإمكان بناء نماذج تعلم آلي تقوم بالمهمة ذاتها: تنطلق من صورة وتقارنها بكافة المعطيات الملتقطة والمخزنة ضمن قاعدة المعطيات.

1.9.2. بيانات تدفقية

أي أنها تعتمد على تسلسل من الصور تمثل مروراً لشخص ما أمام كاميرا ما في لحظة ما سنسمي هذا الشكل من المعطيات أو المدخلات "مرور" وهي في الأدبيات مذكورة تحت اصطلاح "tracklet" أي وحدة بيانات خاصة بالملاحقة.

- ضمن هذا النمط من المعطيات يمكننا التقاط أنماط تعبر عن تغييرات ضمن تسلسل الصور هذا وقد اعتمدنا توافقاً مع الأدبيات ومع الحدس نوعين من هذه المعطيات:
- المشية Gait: وتعبر عن شكل وايقاع تغير مواضع أجزاء الجسم بالنسبة لبعضها البعض خلال طور مشي كامل (خطوتان متتاليتان)
- الهيئة العامة أو Silhouette

الشكل العام للجسم وتغيراته وتغطي هذه المعطيات الاختلافات في حجم الشخص (هنا يمكننا التمييز بين الشخص البدين والشخص العادي مثلاً) أو الشخص الذي يرتدي ملابس سميكة عن غيره.

1.9.3. سمات دلالية ولونية

كأن نستطيع توصيف الشخص من ناحية الجنس، العمر، نمط اللباس الذي يرتديه، الأغراض والاكسسوارات التي يحملها أو يرتديها كحقيبة ظهر أو مظلة مثلاً.

1.10 المساهمات الأساسية للبحث

تتمثّل المساهمات العلمية الأساسية لهذا البحث فيما يلي:

إجراء مراجعة منهجية موسّعة وشاملة لأحدث ما نُشر في مجال أنظمة إعادة التعرّف على الأشخاص (Person Re-Identification)، مع التركيز على الخصائص المميِّزة المستخدمة في النماذج الحديثة، بما يشمل الخصائص المعتمدة على المظهر البصري، والحركة، والوضعيات الجسدية (Pose/Gait)، والتمثيلات الهيكلية، والخصائص الدلالية المستخلصة بواسطة نماذج الذكاء الاصطناعي المتقدمة.

تحليل هذه الخصائص بصورة نقدية وسبرها من حيث القدرة التمييزية، والتعقيد الحسابي، وقابلية التعميم عبر بيئات تصوير مختلفة، إضافةً إلى المقارنة فيما بينها استنادًا إلى نتائج منشورة على مجموعات بيانات قياسية في مجال المراقبة بالفيديو.

دراسة حدود تطبيق مقاربات إعادة التعرّف المختلفة ضمن بيئات الحوسبة الطرفية (Edge Computing)، مع الأخذ بعين الاعتبار القيود العملية المرتبطة بقدرات المعالجة، والذاكرة، واستهلاك الطاقة، ومتطلبات الزمن الحقيقي في أنظمة المراقبة الواقعية.

اقتراح مجموعة من التمثيلات والخصائص الأكثر ملاءمة للتشغيل على الأنظمة الطرفية، بما يوازن بين دقة إعادة التعرّف وكفاءة الموارد، ويأخذ في الحسبان طبيعة البيانات المستمدة من بيئات مراقبة حقيقية وغير مضبوطة.

استكشاف ودراسة طرائق دمج (Fusion) مخرجات عدة نماذج ذكاء اصطناعي متعددة الأنماط، وتحليل أثر هذا الدمج على تحسين متانة النظام ودقته في مواجهة التحديات الواقعية مثل تغيّر الإضاءة، وزوايا التصوير، والانسدالات الجزئية، وتباين أنماط المشي.

تقديم إطار تحليلي يربط بين اختيار الخصائص المميِّزة، واستراتيجية الدمج متعددة النماذج، ومتطلبات النشر العملي لأنظمة إعادة التعرّف في سيناريوهات المراقبة المعتمدة على الحوسبة الطرفية.

1.11 بنية الأطروحة

تتبع الرسالة التي يتم تطويرها هيكلاً منهجياً يهدف إلى معالجة مشكلة إعادة التعرف على الأشخاص (Person Re-ID) في بيئات المراقبة الذكية الموزعة (Big Data & Edge Computing) من خلال دمج خصائص بيومترية متعددة. ينقسم هيكل الرسالة إلى ستة فصول رئيسية بالإضافة إلى الأجزاء التمهيدية والختامية.

الجزء التمهيدي يتولى مهمة تقديم سياق الرسالة، بما في ذلك العنوان الرسمي، صفحة الإهداء، والملخص، وتحديد أسماء الباحث والمشرفين.

الفصل الأول: مقدمة عامة، هو الفصل التأسيسي الذي يحدد المشكلة والدوافع. يبدأ بإطار عام حول أهمية الأمن وأنظمة المراقبة (CCTV) وتحولها نحو الأنظمة الذكية الموزعة التي تولد "المعطيات الكبيرة" (Big Data). يُعرِّف هذا الفصل بوضوح المشكلة العلمية المتمثلة في إعادة المطابقة (Re-identification) والتحديات المرتبطة بها (مثل عدم التعاون، زوايا الكاميرا، وجودة الصور)، ويحدد بوضوح دوافع البحث وأهميته والمساهمات الرئيسية التي يقدمها.

الفصل الثاني: الدراسة النظرية، يعمل على بناء الأساس المعرفي. يقدم المفاهيم والمصطلحات الأساسية اللازمة لفهم المقاربة المقترحة والحلول المعاصرة. يشمل هذا الفصل مقدمات نظرية تشرح المفردات التي سترد لاحقاً في الدراسة المرجعية والحلول المقترحة، يمكن للقارئ تجاوزه والعودة إليه فقط في حال ورود مصطلح غير مألوف بالنسبة إليه.

الفصل الثالث: الدراسة المرجعية، هو فصل مراجعة الأدبيات وتحديد المعايير. يقدم مراجعة منهجية موسّعة للدراسات المنشورة في مجال Person Re-ID. يركز بشكل خاص على تصنيف هذه الدراسات وفقاً للخصائص المميزة المعتمدة (كالمظهر، المشية، والسمات الدلالية). كما يقدم شرحاً مفصلاً لمجموعات البيانات المعيارية المستخدمة في المجال مثل (Market-1501 وDukeMTMC-reID)، ويوضح خصائصها لتدريب وتقييم النماذج المقترحة.

الفصل الرابع: النماذج المقترحة، هو الفصل العملي الذي يقدم فيه الباحث الحلول ومنهجية التنفيذ. يقدم هذا الفصل النماذج المقترحة التي تعتمد على دمج السمات المختلفة (المظهر، الحركة الهيكلية، الظل، السمات الدلالية). يفصّل فيه اختيار الأدوات المناسبة مثل اختيار (YOLOv8-Pose) لدقته وسرعته الملائمة للحوسبة الطرفية ويشرح منهجية استخراج كل سمة على حدة، بما في ذلك استخراج الوضعيات وتجزئة الصور الظلية.

الفصل الخامس: النتائج العملية، مخصص للتقييم الكمي وإظهار التأثير. يعرض هذا الفصل نتائج التجارب العملية التي تم إجراؤها. يشمل وصف بيئة العمل ومجموعات البيانات المستخدمة في التدريب والاختبار. الهدف الأساسي هو قياس أداء النماذج المقترحة، خاصة نماذج الدمج، ومقارنتها بنتائج الأدبيات الحديثة (State-of-the-Art). يحلل الفصل أيضاً تأثير مختلف البنى والتعديلات على دقة وأداء النظام ضمن قيود الحوسبة الطرفية.

الفصل السادس: الخاتمة والآفاق المستقبلية، هو الفصل التلخيصي والتوجيهي. يلخص النتائج والمساهمات الرئيسية التي حققها البحث، ويؤكد على تحقيق أهداف الأطروحة. كما يقترح توجيهات للبحث المستقبلي (Future Work) من خلال تحديد المجالات التي يمكن فيها تحسين النموذج أو استكشاف تحديات جديدة.

أخيراً، يتضمن الجزء الختامي قائمة المراجع (References)، وهي سرد منهجي لجميع المصادر العلمية التي تم الاستناد إليها في البحث.

الفصل الثاني: الدراسة النظرية

1.2. مقدمة

سنستعرض في هذا الفصل العديد من المفاهيم النظرية الضرورية لفهم سياق المسائل المرتبطة بهذا البحث سواء فيما ما يخص الدراسة المرجعية أو الحل المقترح

2.2. نماذج استخراج الوضعيات Pose Extraction Models

2.2.1. وضعية الجسم كبنية معطيات

1. نموذج [1] High-Resolution Network

HRNet (الشبكة عالية الدقة) هي بنية شبكة عصبونية (Neural Network) تم تصميمها خصيصاً للمهام التي تتطلب دقة مكانية عالية، وعلى رأسها تقدير الوضع البشري.

المفهوم الأساسي:

الفكرة الجوهرية لـ HRNet هي الحفاظ على تمثيل عالي الدقة (high-resolution) طوال عملية المعالجة بأكملها.

- البنية التقليدية مثل (ResNet): تبدأ بدقة عالية ثم تقللها تدريجياً (downsampling) لاستخلاص الميزات الدلالية (semantic features)، وفي النهاية تعيد تكبيرها (upsampling) لاستعادة الدقة المكانية، وهذا يؤدي غالباً إلى فقدان التفاصيل الدقيقة.
- **بنية HRNet**:

1. تبدأ بفرع واحد عالي الدقة.
2. تضيف فروعاً منخفضة الدقة بشكل متوازٍ (parallel) في مراحل متتالية.
3. الأهم من ذلك، أنها تحافظ على الفرع الأصلي عالي الدقة وتجري عمليات "دمج" (fusion) متكررة للمعلومات بين جميع الفروع (عالية ومنخفضة الدقة).

طريقة العمل (Top-Down):

تتبع HRNet منهجية "من الأعلى إلى الأسفل" (Top-Down):

1. اكتشاف الشخص: تحتاج أولاً إلى نموذج آخر مثل (YOLO أو Faster R-CNN) لاكتشاف كل شخص في الصورة وتحديد "صندوق محيط" (bounding box) حوله.
2. تقدير الوضع: تقوم HRNet بعد ذلك بتحليل كل صندوق محيط على حدة لتقدير النقاط الرئيسية (keypoints) بداخل ذلك الصندوق فقط.

نقاط القوة والضعف:

- القوة: دقة عالية جداً. نظراً لاحتفاظها بالتفاصيل المكانية الدقيقة طوال الوقت، تُعتبر HRNet واحدة من أدق النماذج لتقدير الوضع البشري وغالباً ما تحقق نتائج متطورة (State-of-the-Art) في معايير الأداء.
- الضعف: بطيئة نسبياً. بنيتها المعقدة التي تحتوي على فروع متوازية وعمليات دمج متكررة تجعلها ثقيلة حسابياً. كما أن اعتمادها على خطوتين (الاكتشاف أولاً ثم التقدير) يزيد من وقت الاستدلال (inference time).

2. نموذج YOLOv8-Pose

YOLOv8-Pose هو ليس بنية جديدة تماماً، بل هو تطبيق متخصص ضمن إطار عمل YOLOv8 (التابع لشركة Ultralytics) ، والذي يدمج مهمة تقدير الوضع مع مهمة اكتشاف الأجسام.

المفهوم الأساسي:

الفكرة الجوهرية لـ YOLOv8-Pose هي السرعة والكفاءة من خلال نموذج شامل (End-to-End).

- البنية: يستخدم بنية 8YOLOv الأساسية (المعروفة بسرعتها الفائقة في اكتشاف الأجسام) ويضيف "رأساً" (head) إضافياً للشبكة.
- رؤوس 8YOLOv:

1. رأس الاكتشاف(Detection Head): يحدد الصناديق المحيطة (bounding boxes) للأشخاص.
2. رأس التجزئة (Segmentation Head): (في نماذج –seg) يحدد أقنعة البكسلات.

3. رأس تقدير الوضع (Pose Head): (في نماذج -pose) يحدد إحداثيات النقاط الرئيسية (مثل المفاصل) للشخص المكتشف.

طريقة العمل (End-to-End):

يتبع YOLOv8-Pose منهجية "شاملة" (End-to-End) أو "أحادية المرحلة" (Single-Stage):

1. خطوة واحدة: يقوم النموذج بتمريرة واحدة (single pass) على الصورة.
2. المخرجات: يُخرج النموذج في نفس الوقت:
   - الصندوق المحيط للشخص.
   - تصنيف الشخص (class).
   - مواقع جميع النقاط الرئيسية (keypoints) المرتبطة بذلك الشخص.

نقاط القوة والضعف:

- **القوة:** سريع جداً .لأنه ينجز الاكتشاف والتقدير في خطوة واحدة، فهو مثالي للتطبيقات التي تتطلب الأداء في الوقت الفعلي (real-time) مثل تحليل الألعاب الرياضية المباشرة أو تطبيقات اللياقة البدنية.
- **القوة:** سهولة الاستخدام: يأتي ضمن إطار عمل متكامل (Ultralytics) يسهل عملية التدريب والنشر.
- **الضعف:** قد يكون أقل دقة (في المطلق) .على الرغم من دقته الممتازة بالنسبة لسرعته، إلا أن النماذج الأثقل والأكثر تعقيداً مثل HRNet (خاصة الإصدارات الكبيرة منها) قد تتفوق عليها في مهام الدقة البحتة (benchmarks).

3. مقارنة مباشرة: HRNet vs. YOLO v8-Pose

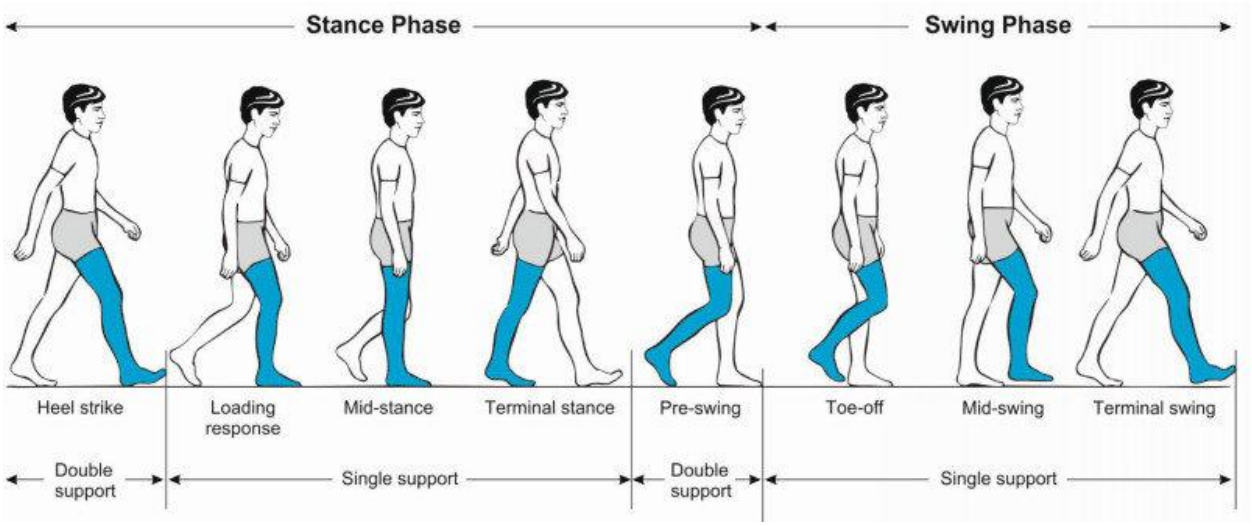
الميزة	ose YOLOv8-P	HRNet (شبكة عالية الدقة)
المنهجية	End-to-End (**شامل / أحادي المرحلة**)	Top-Down (من الأعلى للأسفل)

خطوات العمل	خطوة واحدة: الاكتشاف والتقدير معاً.	خطوتين: 1. اكتشاف الشخص (بنموذج آخر). 2. تقدير الوضع.
السرعة	سريعة جداً (مثالية للوقت الفعلي).	بطيئة نسبياً.
الدقة	دقة جيدة جداً (أفضل توازن بين السرعة والدقة).	عالية جداً (غالباً الأعلى في الدقة البحتة).
التعقيد	خفيفة الوزن وسهلة النشر (خاصة الإصدارات n/s).	معقدة حسابياً وثقيلة.
الاستخدام المثالي	تطبيقات الوقت الفعلي (Real-time)، تتبع الأشخاص، الروبوتات، تطبيقات الويب والهواتف.	التحليل الدقيق بعد التسجيل (Offline analysis)، الأبحاث، عندما تكون الدقة هي الأولوية القصوى.

الجدول 3 مقارنة بين HRENET و YOLO-POSE

2.3. نمط المشي Gait

هو الطريقة المميزة التي يتحرك بها الإنسان أثناء المشي، ويُعد عملية حركية دورية ومعقّدة ناتجة عن تنسيق دقيق بين الجهاز العصبي المركزي والعضلات والعظام والمفاصل. ويعكس نمط المشي خصائص فردية مثل الطول والوزن والعمر والحالة الصحية، كما يتأثر بالعوامل البيئية وسرعة الحركة، ولهذا يُستخدم في الدراسات الطبية والرياضية وأنظمة التعرف الحيوي.



الشكل 2 أطوار عملية المشي عند الإنسان

يمر نمط المشي بدورة تُسمّى دورة المشي **(Gait Cycle)** انظر الشكل، وتبدأ من لحظة ملامسة قدمٍ ما للأرض حتى تلامس القدم نفسها الأرض مرة أخرى. وتنقسم هذه الدورة إلى طورين رئيسيين:

1. **طور الارتكاز (Stance Phase)**

يشكّل نحو 60٪ من دورة المشي، حيث تكون القدم ملامسة للأرض وتتحمّل وزن الجسم. يبدأ هذا الطور بملامسة الكعب للأرض (Heel Strike)، ثم انتقال كامل القدم إلى الأرض، وينتهي بدفع مقدّمة القدم (Toe Off) استعدادًا للحركة التالية.

2. **طور التأرجح (Swing Phase)**

يشكّل نحو 40٪ من دورة المشي، حيث تكون القدم مرفوعة عن الأرض وتتحرك للأمام. يبدأ بعد رفع القدم من الأرض، ويمر بمرحلة تسارع الساق للأمام، ثم يتباطأ قبل أن تلامس القدم الأرض مرة أخرى لبدء دورة جديدة.

يُعد تحليل هذه الأطوار أساسًا لفهم الحركة البشرية، وتشخيص اضطرابات المشي، وكذلك لتطوير أنظمة ذكية تعتمد على نمط المشي كخاصية مميزة للتعرّف على الأشخاص في بيئات المراقبة.

2.4. الشبكات العصبونية البيانية (Graph Neural Networks)[2]

هي فئة متقدمة من نماذج التعلم العميق صُمِّمت خصيصًا لمعالجة البيانات الممثَّلة على شكل رسوم بيانية (Graphs) حيث تتكوّن البيانات من عُقد (Nodes) وروابط فيما بينها تُسمّى حواف (Edges)، وقد تكون لكل عقدة أو حافة خصائص وسمات خاصة. على عكس الشبكات العصبونية التقليدية التي تفترض بنية منتظمة للبيانات (مثل الصور أو السلاسل الزمنية)، تتميز GNNs بقدرتها على العمل مع بُنى غير منتظمة ومعقّدة تعكس العلاقات الحقيقية بين الكيانات.

تعتمد آلية عمل الشبكات العصبونية البيانية على مبدأ تمرير الرسائل (Message Passing)، حيث تقوم كل عقدة بتجميع المعلومات من جيرانها في الرسم البياني، ثم دمج هذه المعلومات مع خصائصها الذاتية لتحديث تمثيلها الداخلي (Embedding). تتكرر هذه العملية عبر عدة طبقات، مما يسمح للنموذج بالتقاط العلاقات المحلية والعالمية داخل الرسم البياني، سواء كانت علاقات مباشرة أو غير مباشرة.

ما يميّز الشبكات العصبونية البيانية هو قدرتها على:

- **نمذجة العلاقات المعقّدة** بين الكيانات بدل الاعتماد على الخصائص الفردية فقط.
- **المرونة العالية** في التعامل مع أحجام ورسوم بيانية مختلفة دون الحاجة لإعادة تصميم النموذج.
- **الاستفادة من البنية الهيكلية للبيانات**، وهو ما يعطيها تفوقًا واضحًا في المسائل التي تكون فيها العلاقات عاملًا حاسمًا.

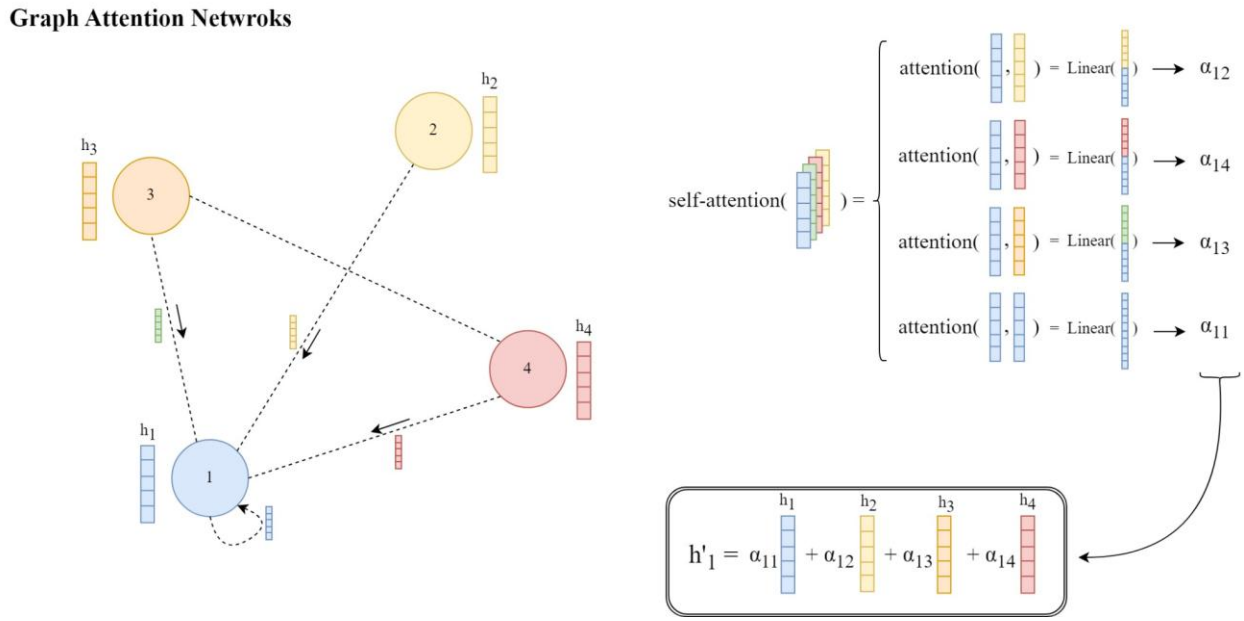
تُستخدم الشبكات العصبونية البيانية في العديد من المجالات، من أبرزها:

- أنظمة التوصية (نمذجة العلاقات بين المستخدمين والعناصر).
- الرؤية الحاسوبية وتحليل الفيديو، مثل تتبع الأجسام أو تمثيل الهيكل العظمي البشري (Skeleton-based Models).
- إعادة التعرّف على الأشخاص (Person Re-Identification)، حيث تُستخدم لتمثيل العلاقات المكانية والزمنية بين أجزاء الجسم أو بين اللقطات المتتابعة.
- الشبكات الاجتماعية لتحليل التأثير والعلاقات.
- الكيمياء والبيولوجيا، مثل نمذجة الجزيئات والبروتينات.

بشكل عام، تُعد الشبكات العصبونية البيانية أداة قوية عندما تكون العلاقات والبنية جزءًا أساسيًا من طبيعة المشكلة، وهو ما يجعلها خيارًا مثاليًا للعديد من التطبيقات الحديثة في الذكاء الاصطناعي وتحليل البيانات المعقّدة.

2.5. شبكات الانتباه البيانية (Graph Attention Networks – GATs)[3]

هي امتداد متقدّم للشبكات العصبونية البيانية، حيث تُدخل آلية الانتباه (Attention Mechanism) ضمن عملية تمرير الرسائل بين العُقد. بدلًا من معاملة جميع العُقد المجاورة بالوزن نفسه أثناء تجميع المعلومات، تقوم GATs بتعلّم أوزان انتباه مختلفة تعبّر عن مدى أهمية كل جار بالنسبة للعقدة المستهدفة. يسمح ذلك للنموذج بالتركيز على العلاقات الأكثر تأثيرًا وتجاهل الضوضاء أو الروابط الأقل صلة.



الشكل 3 شبكات الانتباه البيانية GRAPH ATTENTION NETWORKS

تعتمد GATs على حساب معاملات انتباه قابلة للتعلّم بين العُقد المتجاورة، ثم استخدام هذه المعاملات لوزن الرسائل المتبادلة قبل دمجها. وغالبًا ما تُستخدم آلية الانتباه متعددة الرؤوس (Multi-Head Attention) لتحسين الاستقرار والقدرة التمثيلية، حيث تُلتقط أنماط علاقات مختلفة بالتوازي داخل الطبقة الواحدة. هذه المقاربة تجعل GATs أكثر تكيفًا ومرونة مقارنةً بـ GNNs التقليدية التي تعتمد على تجميع ثابت أو متوسط بسيط.

تتميّز شبكات الانتباه البيانية بما يلي:

- تمييز الأهمية النسبية للعلاقات بين العُقد بدل افتراض تساويها.
- تحسين الأداء في الرسوم البيانية غير المتجانسة أو المليئة بالضوضاء.

- قابلية تفسير أفضل نسبيًا، إذ يمكن تحليل أوزان الانتباه لفهم سبب تأثير عقد معيّنة أكثر من غيرها.

تُستخدم GATs بكفاءة في مهام مثل تصنيف العُقد، التنبؤ بالروابط، تحليل الشبكات الاجتماعية، وكذلك في تطبيقات الرؤية الحاسوبية المعتمدة على الرسوم البيانية، مثل نمذجة العلاقات بين أجزاء الجسم في تتبّع الحركة والمشية (Gait Analysis) أو تمثيل العلاقات الزمنية والمكانية في مقاطع الفيديو. لذلك تُعد GATs خيارًا قويًا عندما تكون أهمية العلاقة نفسها متغيرة وتعتمد على السياق، لا مجرد وجود الاتصال بين العُقد.

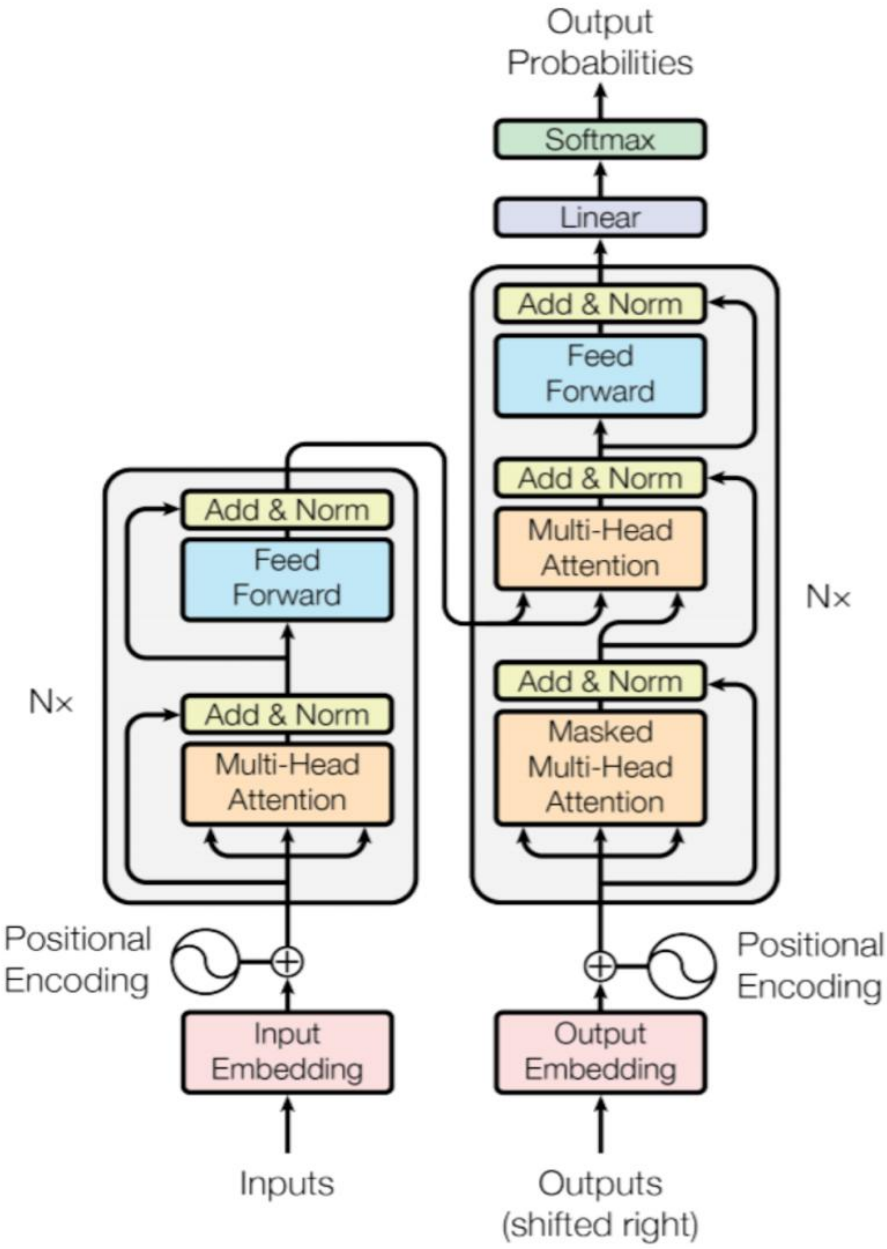
2.6. المحولات [4] Transformers

يُقصد باصطلاح Transformer أنه بنية معمارية (Architecture) للشبكات العصبية العميقة صُمِّمت لمعالجة البيانات المتسلسلة عبر نمذجة العلاقات بين عناصر الدخل باستخدام آلية الانتباه الذاتي (Self-Attention) بدلًا من الاعتماد على الترتيب الزمني أو التكرار كما في الشبكات المتكررة التقليدية.

يعني مصطلح Transformer حرفيًا «المحوِّل»، أي النموذج القادر على تحويل تمثيل الدخل من فضاء إلى آخر أكثر دلالة عبر حساب مدى أهمية كل عنصر بالنسبة لبقية العناصر داخل التسلسل نفسه. وبذلك، فإن الـ Transformer لا ''يقرأ'' البيانات خطوة بخطوة، بل ينظر إليها كوحدة مترابطة، ويُعيد وزن عناصرها بحسب علاقاتها السياقية، مما يسمح بفهم أعمق للاعتماديات القريبة والبعيدة على حدّ سواء.

وبصيغة مختصرة يمكن القول:

الـ Transformer هو إطار شبكي يعتمد على الانتباه الذاتي لتعلّم تمثيلات سياقية غنية للبيانات المتسلسلة، مع قابلية عالية للمعالجة المتوازية والكفاءة الحسابية.



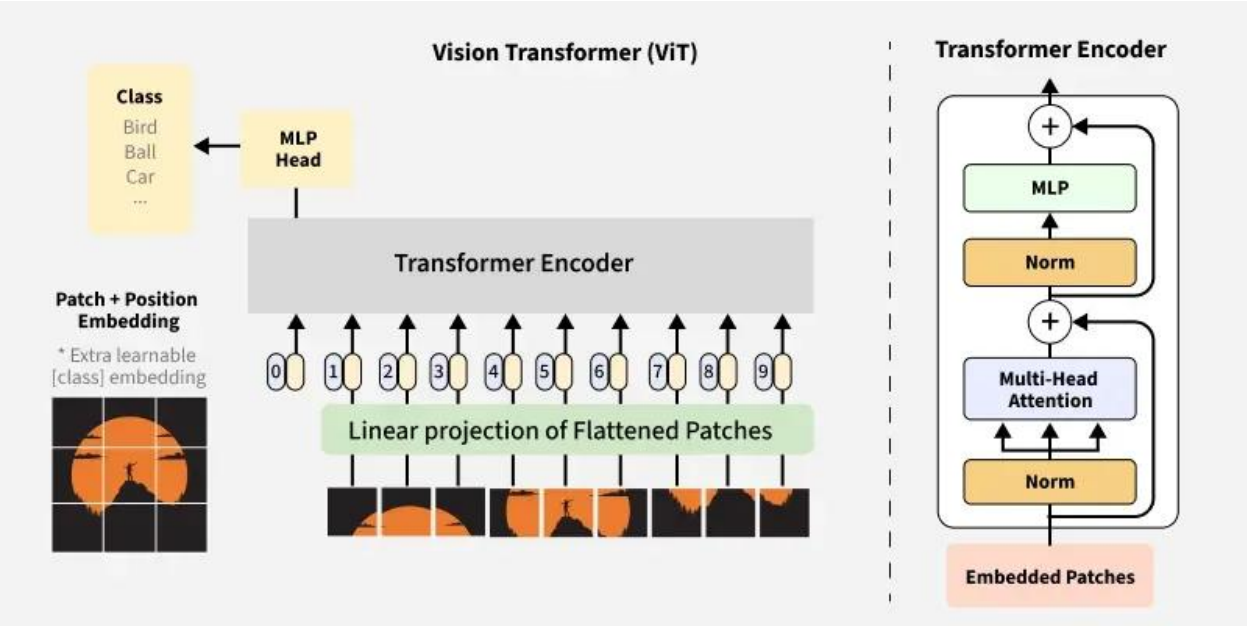
الشكل 4 معمارية المحولات TRANSFORMERS

تُعد الشبكات العصبونية المبنية على بنية Transformers من أهم التطورات الحديثة في مجال الذكاء الاصطناعي والتعلّم العميق، إذ تعتمد بشكل أساسي على آلية الانتباه الذاتي (Self-Attention) بدلًا من التتابعية الزمنية المستخدمة في الشبكات المتكررة التقليدية (RNN وLSTM). تتيح هذه البنية للنموذج فهم العلاقات بين جميع عناصر الدخل في الوقت نفسه، مما يمكّنه من التقاط الاعتمادات بعيدة المدى بكفاءة عالية. تتكون بنية الـ Transformer عادةً من وحدات Encoder وDecoder، حيث تتضمن كل وحدة طبقات انتباه متعددة الرؤوس (Multi-Head Attention) وطبقات تغذية أمامية (Feed-

Forward Networks)، إضافةً إلى آليات التطبيع (Layer Normalization) والروابط المتبقية (Residual Connections). ما يميز هذه الشبكات هو قابليتها العالية للتوازي أثناء التدريب، ودقتها الكبيرة في تمثيل السياق، الأمر الذي جعلها الأساس للعديد من النماذج المتقدمة في معالجة اللغة الطبيعية، والرؤية الحاسوبية، والأنظمة متعددة الوسائط، بل وحتى في تطبيقات متقدمة مثل إعادة التعرّف (Re-Identification) وتحليل التسلسلات الزمنية المعقّدة.

2.7. الشبكات العصبونية البصرية القائمة على بنية المحوِّل [5](Vision Transformer)

هي امتداد مباشر لفلسفة نماذج الـ Transformer إلى مجال الرؤية الحاسوبية، حيث يتم التعامل مع الصورة على أنها تسلسل من الرقع (Patches) بدلًا من معالجتها عبر المرشحات التلافيفية التقليدية. في هذه البنية، تُقسَّم الصورة إلى رقع ثابتة الحجم، ثم يُحوَّل كل منها إلى تمثيل متجهي يُغذّى إلى نموذج Transformer باستخدام آلية الانتباه الذاتي، مما يمكّن النموذج من تعلّم العلاقات المكانية والسياقية بين مختلف أجزاء الصورة بشكل عالمي. وعلى عكس الشبكات التلافيفية (CNNs) التي تعتمد على مجال رؤية محلي، يسمح ViT بنمذجة الاعتمادات بعيدة المدى بين الرقع منذ المراحل الأولى، الأمر الذي يعزز قدرته على تمثيل البنية الشمولية للمشهد البصري. وتتكون بنية Vision Transformer عادةً من طبقة تضمين للرقع (Patch Embedding)، تليها سلسلة من طبقات Transformer Encoder، مع إضافة تمثيل موضعي (Positional Encoding) للحفاظ على المعلومات المكانية. وقد أثبتت هذه البنية فعاليتها العالية في العديد من مهام الرؤية الحاسوبية، مثل تصنيف الصور، واكتشاف الأجسام، وإعادة التعرّف (Re-Identification)، خاصةً عند توفر بيانات تدريب كبيرة أو عند دمجها مع استراتيجيات تدريب مسبق متقدمة.

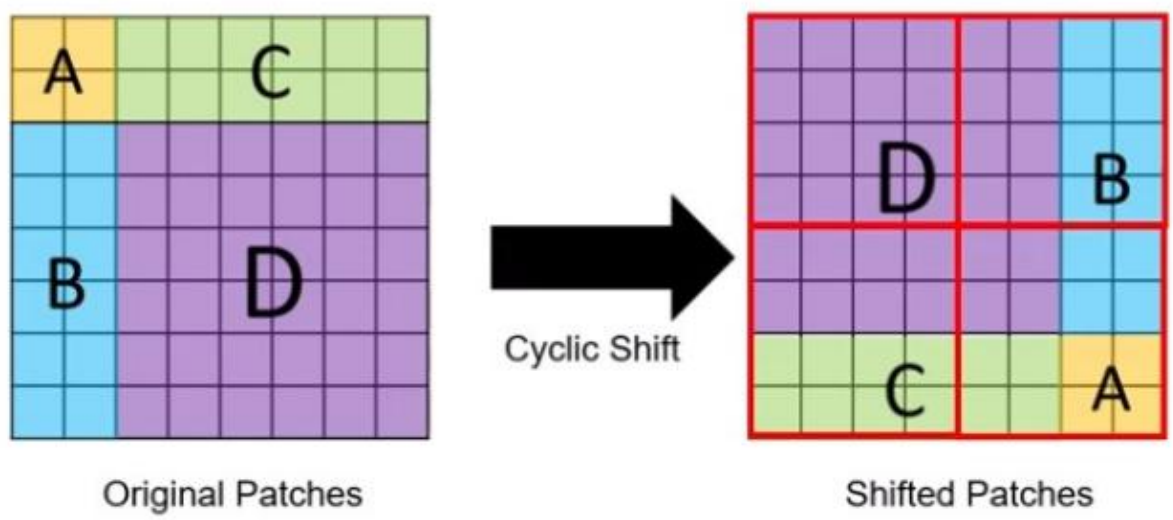


الشكل 5 بنية المحول البصري VISION TRANSFORMER VIT

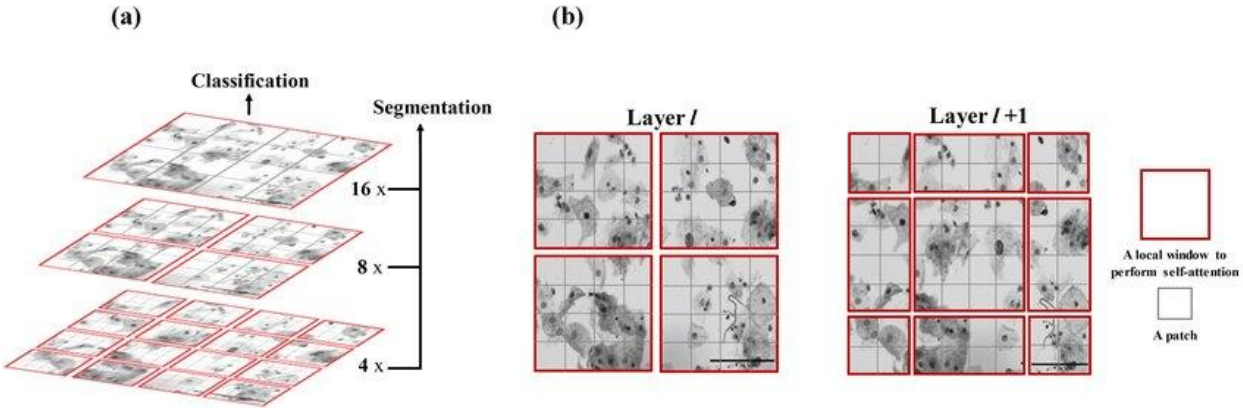
2.8. محوِّل النوافذ المُزاحة [6] (Shifted Window Transformer – Swin)

يُعد Swin Transformer (اختصارًا لـ Shifted Window Transformer) تطويرًا متقدمًا لبنية المحوِّلات البصرية، صُمِّم خصيصًا لمعالجة القيود الحسابية التي تواجه نماذج Vision Transformer التقليدية عند التعامل مع الصور عالية الدقة.

تعتمد هذه البنية على آلية الانتباه الذاتي ضمن نوافذ محلية (Window-based Self-Attention) بدلًا من الانتباه العالمي، مما يقلّل التعقيد الحسابي بشكل ملحوظ. ولضمان تبادل المعلومات بين النوافذ المختلفة، يستخدم Swin مفهوم النوافذ المنزاحة (Shifted Windows)، حيث يتغير تموضع النوافذ بين الطبقات المتتالية، الأمر الذي يسمح للنموذج بالتقاط العلاقات المكانية عبر حدود النوافذ دون كلفة حسابية عالية. كما يتميز Swin ببنية هرمية متعددة المراحل مشابهة للشبكات الالتفافية، حيث تتغير أبعاد التمثيل تدريجيًا مع زيادة العمق، مما يجعله ملائمًا لمهام الرؤية الحاسوبية متعدّدة المقاييس.



الشكل 6 اعادة ترتيب النوافذ في خوارزمية SWIN

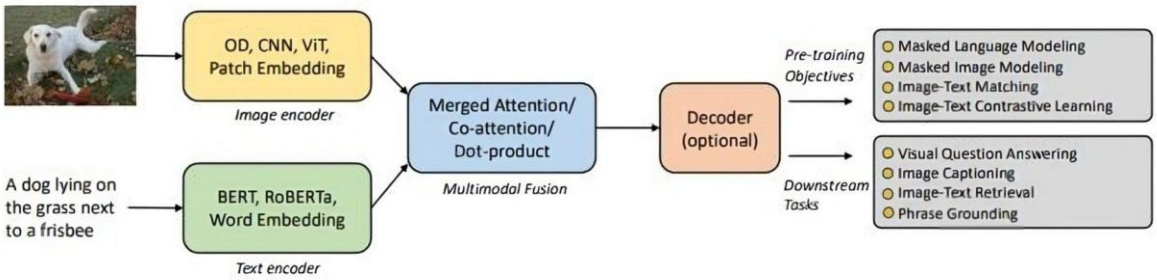


الشكل 7 البنية الهرمية في SWIN

وقد أثبت Swin Transformer كفاءة عالية في تحقيق توازن بين تمثيل العلاقات العالمية والكفاءة الحسابية، مما جعله خيارًا شائعًا في تطبيقات متقدمة مثل تصنيف الصور، واكتشاف الأجسام، والتجزئة الدلالية، وأنظمة إعادة التعرّف (Re-Identification)، خاصةً في البيئات الواقعية ذات القيود الحوسبية.

2.9. نماذج اللغة الكبيرة ونماذج الرؤية–اللغة LLMs and VLMs

تُعد نماذج اللغة الكبيرة (Large Language Models – LLMs) من أبرز إنجازات الذكاء الصنعي المعاصر، إذ تعتمد على بنية المحوِّل (Transformer) لتعلّم تمثيلات لغوية عميقة من كميات هائلة من البيانات النصية، بما يمكّنها من فهم السياق، واستنتاج المعاني، وتوليد نصوص مترابطة ودقيقة على مستوى دلالي عالٍ. وتتميّز هذه النماذج بقدرتها على نمذجة العلاقات المعقّدة بين الكلمات والجمل عبر آلية الانتباه الذاتي، مما يسمح لها بالتعامل مع مهام لغوية متقدمة مثل التلخيص، والترجمة، والإجابة عن الأسئلة، والاستدلال المنطقي.



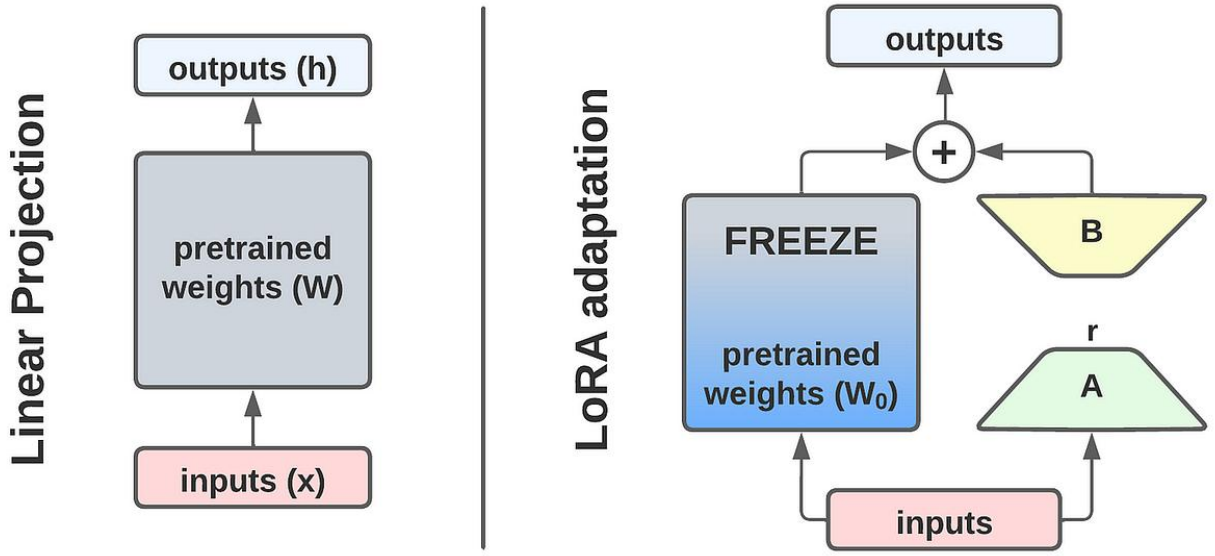
الشكل 8 نموذج لغوي بصري

وانطلاقًا من هذا التطور، ظهرت نماذج الرؤية–اللغة (Vision-Language Models – VLMs) بوصفها امتدادًا طبيعيًا لنماذج اللغة الكبيرة نحو المجال متعدد الوسائط، حيث تهدف إلى دمج المعلومات البصرية والنصية ضمن فضاء تمثيلي موحّد. تعتمد هذه النماذج على مواءمة مُشفّر بصري مثل (Vision Transformer) مع مُشفّر لغوي قوي قائم على LLM، مما يتيح الربط الدلالي بين العناصر المرئية ومقابلاتها اللغوية. ويُكسب هذا الدمج نماذج VLM قدرةً متقدمة على فهم المشاهد البصرية في سياق لغوي غني، وتنفيذ مهام معقّدة مثل وصف الصور، والإجابة عن الأسئلة البصرية، والاسترجاع متعدد الوسائط، واستخلاص السمات الدلالية عالية المستوى. وبذلك، تمثل نماذج LLM وVLM معًا نقلة نوعية من المعالجة الأحادية إلى الفهم السياقي الشامل، وهو ما يجعلها ركيزة أساسية في الأنظمة الذكية الحديثة والتطبيقات البحثية المتقدمة.

2.10. المعايرة الدقيقة للنماذج اللغوية الكبيرة Fine Tuning LLMs

يُشير اصطلاح المعايرة الدقيقة (Fine-Tuning) لنماذج اللغة الكبيرة إلى عملية تكييف نموذج لغوي مُدرَّب مسبقًا على نطاق واسع ليؤدي مهامًا أو يعمل ضمن نطاقات تطبيقية محددة، وذلك عبر إعادة تدريبه جزئيًا باستخدام بيانات متخصصة أصغر حجمًا وأكثر تركيزًا. تهدف هذه العملية إلى الاستفادة من المعرفة العامة التي اكتسبها النموذج خلال التدريب المسبق، مع تحسين أدائه في سياق معيّن مثل مجال طبي، قانوني، تقني، أو متعدد الوسائط. ويُعد الضبط الدقيق عنصرًا محوريًا في تحويل نماذج LLM من أدوات عامة إلى مكونات فعّالة داخل أنظمة ذكية تطبيقية، مع تقليل كلفة التدريب مقارنة بإعادة التدريب من الصفر.

ومن بين أكثر تقنيات المعايرة الدقيقة كفاءةً برزت خوارزمية LoRA (Low-Rank Adaptation) [7]، التي تقوم على مبدأ تجميد أوزان النموذج الأساسي وإضافة مصفوفات منخفضة الرتبة إلى طبقات محددة (غالبًا طبقات الإسقاط في آلية الانتباه الذاتي)، بحيث يتم تعلّم هذه الإضافات فقط أثناء التدريب. يسمح هذا النهج بتقليل عدد المعاملات القابلة للتعلّم بشكل كبير، مع الحفاظ على أداء قريب من الضبط الكامل للنموذج. وقد تطورت LoRA إلى عدة أشكال، منها LoRA القياسية، وQLoRA التي تجمع بين LoRA والتكميم منخفض الدقة لتقليل استهلاك الذاكرة، إضافةً إلى متغيرات تستهدف طبقات أو رؤوس انتباه محددة، مما يوفّر مرونة عالية في موازنة الأداء مع القيود الحوسبية.



الشكل 9 مبدأ عمل خوارزمية LoRA

2.11. الحوسبة الحافية Edge Computing

تُعرَّف الحوسبة الحافّية (Edge Computing) بأنها نموذج حوسبي يهدف إلى نقل المعالجة والتخزين واتخاذ القرار من مراكز البيانات السحابية البعيدة إلى أطراف الشبكة، أي بالقرب من مصادر توليد البيانات مثل الحساسات، الكاميرات، والأجهزة الذكية. يكتسب هذا النموذج أهمية متزايدة مع التطور السريع في الذكاء الصنعي (Artificial Intelligence)، إذ تتطلب تطبيقاته الحديثة—ولا سيما تلك المعتمدة على الرؤية الحاسوبية، والتعلّم العميق، وإنترنت الأشياء—زمن استجابة منخفضاً، واعتمادية عالية، وتقليلاً للاعتماد على الاتصال الدائم بالسحابة. يتيح دمج الذكاء الصنعي ضمن الحوسبة الحافية تنفيذ نماذج التحليل والاستدلال محلياً على الأجهزة الطرفية، مثل اكتشاف الأحداث، والتعرّف على الأنماط، واتخاذ قرارات آنية دون الحاجة إلى إرسال كميات ضخمة من البيانات الخام عبر الشبكة، مما يقلل من زمن التأخير (Latency) واستهلاك النطاق الترددي، ويعزز الخصوصية والأمان. ومع ذلك، يفرض هذا التكامل تحديات تقنية مهمة، أبرزها محدودية الموارد الحاسوبية والطاقة في بيئات الحافة، ما يستلزم تطوير نماذج ذكاء صنعي خفيفة الوزن، وتقنيات ضغط وتسريع، وأساليب نشر وتحديث فعّالة. وبناءً عليه، تُعد الحوسبة الحافية المدعومة بالذكاء الصنعي اتجاهاً محورياً في بناء أنظمة ذكية قادرة على العمل في الزمن الحقيقي ضمن بيئات واقعية ومعقدة، مثل أنظمة المراقبة، المركبات الذكية، الرعاية الصحية، والصناعة الذكية، خاصة في السياقات التي تتطلب استقلالية وموثوقية عالية.

تعود الأهمية المتزايدة للحوسبة الحافّية المدعومة بالذكاء الصنعي إلى مجموعة عوامل تقنية وتطبيقية متداخلة، من أبرزها:

1. **الانفجار في حجم البيانات**

الانتشار الواسع لأجهزة إنترنت الأشياء والكاميرات والحساسات يولّد كميات هائلة من البيانات المستمرة، ويصبح إرسالها بالكامل إلى السحابة غير عملي من حيث الكلفة والزمن، ما يجعل المعالجة القريبة من المصدر ضرورة.

2. **الحاجة إلى الاستجابة الآنية (Low Latency)**

العديد من التطبيقات الحديثة—مثل أنظمة المراقبة الذكية، المركبات الذاتية، والأنظمة الصناعية—تتطلب قرارات فورية لا تحتمل تأخير الشبكة، وهو ما توفره المعالجة على الحافة.

3. **تقليل الاعتماد على الاتصال بالسحابة**

في البيئات غير المستقرة شبكياً أو المعزولة (مصانع، حدود، مواقع نائية)، تتيح الحوسبة الحافّية استمرارية العمل حتى مع انقطاع الاتصال أو ضعفه.

4. **تعزيز الخصوصية والأمان**

معالجة البيانات الحساسة محلياً (مثل الوجوه أو السجلات الصحية) تقلل من نقل البيانات الخام عبر الشبكات، ما يحد من مخاطر الاختراق ويلبي متطلبات الامتثال والخصوصية.

5. **خفض كلفة النطاق الترددي والبنية التحتية**

بدلاً من بث البيانات الخام، تُرسل فقط النتائج أو الملخصات إلى السحابة، مما يقلل استهلاك النطاق الترددي وتكلفة التخزين والمعالجة المركزية.

6. **نضج تقنيات الذكاء الصنعي الخفيفة**

تطور نماذج مضغوطة، وتسريع العتاد (مثل وحدات المعالجة العصبية)، وتقنيات التحسين جعل تشغيل الذكاء الصنعي على أجهزة محدودة الموارد أمراً عملياً وفعالاً.

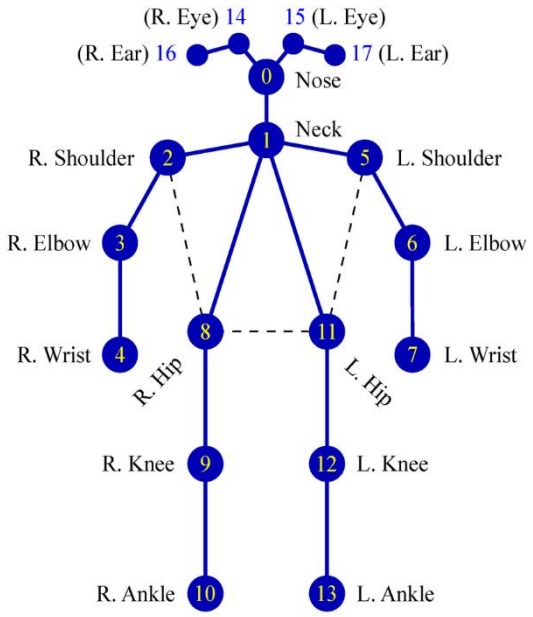
7. **التوسع في التطبيقات الحرجة**

قطاعات مثل الصناعة الذكية، الصحة، الأمن، والمدن الذكية تتطلب موثوقية عالية واستقلالية تشغيلية، وهو ما توفره أنظمة الحافة الذكية.

بناءً على ذلك، لم تعد الحوسبة الحافّية خياراً تكميلياً، بل أصبحت ضرورة معمارية لتمكين الذكاء الصنعي من العمل بكفاءة وموثوقية في العالم الحقيقي، خاصة في التطبيقات الزمنية والحرجة.

2.12. تمثيل COCO-17 لمفاصل جسم الإنسان في تقدير الوضعية البشرية [8]

يُعدّ **تمثيل** COCO-17 أحد أكثر التمثيلات القياسية انتشارًا في مجال تقدير وضعية جسم الإنسان (Human Pose Estimation)، حيث يصف الجسم البشري عبر 17 نقطة مفصلية ثنائية الأبعاد تغطي الرأس، الجذع، والأطراف العلوية والسفلية. يشمل هذا التمثيل نقاطًا مثل الأنف، العينين، الأذنين، الكتفين، المرفقين، الرسغين، الوركين، الركبتين، والكاحلين، مع تعريف ترابطي (Skeleton Graph) يحدد الاتصالات التشريحية بينها.



الشكل 10 تمثيل Coco-17 لوضعية جسم الإنسان

يوفّر هذا البناء توازنًا عمليًا بين الدقة الهندسية والكلفة الحسابية، ما يجعله مناسبًا للتطبيقات الزمن–حقيقية والأنظمة المضمنة، وكذلك للتعلّم العميق على مجموعات بيانات واسعة. وبفضل اعتماده الواسع في COCO Dataset، أصبح COCO-17 معيارًا مرجعيًا للتوافق والمقارنة بين النماذج، كما يُستخدم بكثرة في مهام لاحقة مثل التعرّف القائم على المشية، تتبّع الأشخاص، وتحليل الحركة، مع قابلية التوسّع إلى تمثيلات ثلاثية الأبعاد عبر الرفع (Lifting)

أو النمذجة الزمنية للتسلسلات.

2.13.خاتمة

بهذا نكون قد استعرضنا شرحاً سريعاً لأغلب مالمصطلحات التي سنتطرق إليها في ما بعد وهو الأمر الضروري لفهم أفضل للفصول القادمة.

الفصل الثالث: الدراسة المرجعية

3.1. مقدمة

سنستعرض في هذا الفصل مجموعة الدراسات التي قمنا بالرجوع إليها مبوبة وفقاً للخصائص المميزة التي تركز عليها سنقوم بتفصيل مجموعات البيانات المعيارية المستخدمة أثناء العمل في هذا البحث، وقد يكون من المفيد البدء بشرح مجموعات البيانات، كبنية واختلافات فيما بينها لتوضيح المسألة قبل البدء بتناول الدراسات المختلفة ومقاربات المسألة من انطلاقاً من خصائص مميزة مختلفة.

3.2 . أنماط مجموعات البيانات المستخدمة في مسألة إعادة المطابقة

تصنف مجموعات البيانات المستخدمة في مسألة إعادة المطابقة ضمن ثلاث أصناف رئيسية:

3.2.1. مجموعات إعادة التعرّف المعتمدة على الصور (Image-Based Appearance Re-ID)

وهي مجموعات بيانات تحتوي صوراً لأشخاص تم التقاطها في ازمنة مختلفة ومن كاميرات مختلفة وهي على الرغم

أمثلة:

- Market-1501
- CUHK03

الخصائص التعريفية

- صور ثابتة مستقلة تمثل التقاطات لأشخاص.
- لا يوجد ترابط زمني صلب.
- مُحسّنة لاستخراج خصائص الملابس، اللون، والملمس (Texture).
- غالبًا ما تكون مربعات الاحتواء مقتطعة بإحكام.

ما الذي تدعمه؟

- تمثيلات المظهر باستخدام CNN أو ViT.
- التمييز الهويّاتي اعتمادًا على إطار واحد.

ما الذي لا تدعمه؟

- نمذجة المشية.
- ديناميكيات الوضعية.
- أي شكل من أشكال الاستدلال ثلاثي الأبعاد أو شبه ثلاثي الأبعاد.

مجموعات إعادة التعرّف المعتمدة على الفيديو (Video-Based Appearance Re-ID) (مظهر زمني، وليس مشية)

أمثلة

- MARS
- DukeMTMC-VideoReID

الخصائص التعريفية

- مقاطع فيديو قصيرة (Tracklets).
- تجميع زمني لخصائص المظهر (RGB Images).
- المشي موجود عرضيًا وليس منظمًا.
- دورات المشية غالبًا غير مكتملة.

ما الذي تدعمه؟

- التجميع الزمني (Temporal Pooling).
- نماذج RNN أو آليات الانتباه المعتمدة على المظهر.
- خطوط أساس لإعادة التعرّف المعتمدة على الفيديو.

ما الذي لا تدعمه؟

- نمذجة المشية الدورية.
- ظلال مستقرة (Silhouettes).
- مسارات وضعية موثوقة.

الزمن موجود، لكن الحركة ليست هي الإشارة.

3.2.2. مجموعات البيانات الموجّهة للمشية (Gait-Oriented Datasets)

(الحركة هي السمة الحيوية)

أمثلة

- CASIA-B
- OU-MVLP
- GREW
- Gait3D
- SUSTech1K

الخصائص التعريفية

- تسلسلات مشي طويلة ومتواصلة.
- دورات مشية كاملة.
- دورية واضحة للحركة.
- غالبًا متعددة الزوايا (360 درجة).
- مصمّمة لضمان استقرار الظلال أو الوضعيات.

ما الذي تدعمه؟

- مجسّمات الظلال (شبه ثلاثية الأبعاد).
- ديناميكيات الهيكل العظمي.
- تعلّم هندسة ثلاثية الأبعاد بشكل ضمني.
- تركز على التقاط الحركة بأطوارها الكاملة.

يلخص الجدول التالي الفروقات الأساسية.

الفئة	وجود الحركة	استخدام الحركة كإشارة	دورات مشي مكتملة	ثلاثي الأبعاد أو شبه ثلاثي الأبعاد
إعادة تعرف بالصور	F	F	F	F
إعادة تعرف بالفيديو	T	F	F	F
بيانات المشية	T	T	T	T

الجدول4 جدول المقارن بين أنماط مجموعات البيانات

على الرغم من أنّ مجموعة بيانات Market-1501 قد تتضمّن عدة صور للشخص نفسه خلال مرور واحد أمام الكاميرا، إلا أنّها لا تحمل دلالة زمنية كافية تجعلها صالحة لنمذجة المشية. فهذه الصور تُخزَّن وتُعالَج كعينات مستقلة من دون ترتيب زمني صريح، ولا تتضمن معلومات عن الاستمرارية أو معدل الالتقاط، كما تفتقر إلى الترابط الحركي اللازم لتتبّع تغيّر وضعية الجسم عبر الزمن. إضافة إلى ذلك، لا تضمن Market-1501 اكتمال دورات مشي أو انتظامًا حركيًا يمكن الاعتماد عليه لاستخلاص سمات مشية مستقرة. ويعود ذلك إلى أنّ تصميمها موجّه أساسًا لاستخراج خصائص المظهر مثل الملابس واللون وليس لتحليل الحركة، حيث يُهمل البعد الزمني بالكامل في التمثيل الخوارزمي. وبناءً عليه، فإن إدخال Market-1501 ضمن مسار المشية يؤدي إلى تعلّم ضجيج بصري بدلًا من إشارات حركية ذات معنى، مما يبرّر استبعادها منهجيًا من مسار نمذجة المشية وحصر استخدامها — عند الحاجة — في مسار المظهر فقط

كذلك الأمر في مجموعات بيانات الفيديو مثل MARS وهي النسخة الفيديوية من Market1501 فهي تحوي على tracklets تتضمن صوراً عديدة متتالية لمرورات لأشخاص إلا أنها لا تضمن دورات مشي كاملة و وفي الكثير من حالاتها لا تحوي عدداً كاف من الأطر لتتضمن خصائص حركة ولا تعرف العلاقة الزمنية بين الأطر سوى بأنها متتالية و كما أنها لا تشمل على مناظير كافية تسمح بفهم ثلاثي الابعاد او شبيه بثلاثي الابعاد وهي مجموعة بيانات ذات ضجيج مرتفع العديد من ال tracklets تحوي حجباً جزئيا لأجسام المارة سواء بعوائق أو بأشخاص آخرين.

3.3. مجموعات البيانات المعيارية المتعلقة بمسألة إعادة المطابقة

3.3.1. صورة عامة عن مجموعات البيانات المعيارية ذات الصلة بمسألة إعادة المطابقة

تتشابه مجموعات البيانات المستخدمة في مجال أبحاث إعادة المطابقة في بنيتها، حيث تعتمد على صور ملتقطة لأشخاص عبر أنظمة مراقبة. تكون هذه الصور ناتجة عن عملية الكشف (كشف المارة pedestrian detection) كما في الشكل:

ترافق كل صورة بيانات وصفية metadata تشمل ما يلي:

- معرّف الشخص ذاته
- معرّف الكاميرا
- معرّف الالتقاط (ترتيب الصورة ضمن سلسلة الصور في نفس الالتقاط)

غالباً ما تكون مضمنة كلها في تسلسل المحارف الذي يشكل اسم ملف الصورة ضمن مجموعة البيانات أو حتى اسم المجلد الفرعي ضمنها.

تشمل مجموعات البيانات عدة صور أو إطارات لمرور الشخص أمام كاميرا ما كما في الشكلين التاليين حيث يمثل كل شكل مروراً للشخص ذاته أمام كاميرتين مختلفتين في وقتين مختلفين.

وهنا يجب التنويه إلى أن عدد الإطارات الملتقط يختلف من قد مجموعة بيانات لأخرى ومن التقاط مرور لمرور آخر، وهنا تقسم مجموعات البيانات إلى صنفين مجموعات صورية ومجموعات فيديوية، حيث أن تلك الأخيرة تحرص على أن تقوم بتعريف معرف لعنصر متابعة مرور (tacklet) وعنصر يعرف الصورة الواحدة ضمن المرور كأن يكون اسم ملف الصورة:

Person_id-camera_id_tracklet_id_sequence_id

وتحرص على أن يكون كل مرور ملتقط مؤلفاً من عدد كاف من الأطر (frames) بحيث يستطيع تغطية دور مشي كامل بطوريه (تقديم الرجل الأولى ثم لحاق الاخرى بها قبل أن تتقدم هي بدورها لخطوة جديدة) تستخدم هذه المجموعات في تدريب النماذج التي تركز على الخصائص المرتبطة بالحركة والمشية كسمة حيوية.

يميز الباحثون بين مجموعات البيانات وفقاً لكونها ملتقطة ضمن بيئة داخلية outdoor أو داخلية indoor حيث أن لكلا الصنفين تحديات واعتبارات مختلفة.

نلاحظ في الشكلين السابقين مدى الاختلاف في منظور الرؤية للشخص ذاته من كاميرتين مختلفتين واختلاف الخصائص المميزة التي يمكن التقاطها.

إن المرور الذي قام به الشخص أمام الكاميرا الاولى وحركته من اليمين إلى اليسار ومواجهة مسقطه الجانبي لعدسة الكاميرا يسمح بالتقاط تغيرات حركية جيدة تعبر عن إيقاع مشيته، على عكس الصور التي تمثل مروره الثاني والتي يكون فيها ظهره مواجهاً للكاميرا الأمر الذي يجعل إدراك شكل مشيته من خلالها أصعب، كذلك الأمر بالنسبة للصور الظلية إذ لا يظهر في السلسلة الثانية أثر لحقيبة الظهر من ناحية بل إن الصورة الظلية للشخص ذاته لو لم يكن يرتدي هذه الحقيبة تكاد تكون مطابقة لصورته الظلية مع الحقيبة من هذه الزاوية، يوضح هذا المثال كيف يمكن أن تختلف تمييزية خاصية ما من حالة لأخرى بشكل كبير في مسألة إعادة المطابقة بسبب طبيعة المعطيات المختلفة.

معظم مجموعات البيانات الخاصة بإعادة المطابقة تتألف من صور بصيغتها الخام RGB وبعضها تكون صورها بصيغة ثنائية (أبيض وأسود) أو صور ظلية فقط.

كما توجد مجموعات بيانات تشمل وصوفات نصية (captions) بالإضافة إلى معرفات الأشخاص والكاميرات وتستخدم هذه المجموعات مع النماذج التي تهتم بالسمات الدلالية والوصفية النصية.

في هذا القسم، سيتم استعراض أهم مجموعات البيانات المستخدمة في هذا المشروع:

ICFG-PEDES، CASIA-B، MARS، DukeMTMC-VideoReID، DukeMTMC-reID، Market-1501، وCUHK03 نسخة (TIP-CB) ، مع توضيح الجهة المطوِّرة، وخصائص كل مجموعة، وأهم سيناريوهات الاستخدام المناسبة لها.

3.3.2. مجموعة بيانات Market-1501[9]

الجهة المطوِّرة:

Liang Zheng وزملاؤه، وقد قُدِّمت في ورقة *Scalable Person Re-identification: A Benchmark* في مؤتمر ICCV 2015. (cv-foundation.org)

نظرة عامة:

تُعد Market-1501 واحدة من أوائل وأشهر مجموعات البيانات واسعة النطاق في إعادة التعرّف على الأشخاص من صور ثابتة. تم جمع البيانات باستخدام ست كاميرات مراقبة في بيئة خارجية داخل حرم جامعي (جامعة تسينغهوا).

الخصائص الرئيسة:

- النوع: صورة ثابتة (Image-based ReID).
- البيئة: خارجي (Outdoor)، حرم جامعي شبه منظَّم.
- طبيعة البيانات: صور RGB مقصوصة لأشخاص.
- عدد الكاميرات: 6.
- الحجم: حوالي 32,668 صورة لـ 1,501 هوية(cv-foundation.org) .
- تم الحصول على مربعات الإحاطة (Bounding Boxes) تلقائيًا باستخدام كاشف DPM، ما يحاكي تطبيقات حقيقية مع نظام كشف آلي.

السمات المميِّزة:

- درجة انسداد منخفضة نسبيًا وانضباط في ظروف الإضاءة.
- تنسيق بسيط وواضح، ما جعلها معيارًا أساسيًا لبناء واختبار نماذج ReID القائمة على الصور فقط.

أفضل حالات الاستخدام:

- تدريب واختبار النماذج الكلاسيكية لـ ReID بالصورة فقط مثل (ResNet، OSNet، TransReID).

- إجراء تجارب أولية سريعة ومقارنة أداء النماذج قبل الانتقال لمجموعات أكثر تعقيدًا.

3.3.3. مجموعة بيانات DukeMTMC-reID [10]

الجهة المطوِّرة:

مشتقة من مجموعة التتبع متعددة الكاميرات DukeMTMC، وقد نُظِّمت بصيغة ReID في ورقة DukeMTMC4ReID: A Large-Scale Multi-Camera Person Re-Identification Dataset بواسطة Mengran Gou وزملائه، في ورشة CVPR 2017. (CVF Open Access)

نظرة عامة:

تم إنشاء DukeMTMC-reID من بيانات تتبع متعددة الأهداف في حرم جامعة Duke، مع التركيز على مهمة إعادة التعرّف من صور ثابتة مع ثمانِ كاميرات مراقبة متزامنة. (CVF Open Access)

الخصائص الرئيسة:

- النوع: صورة ثابتة (Image-based ReID).
- البيئة: خارجي (Outdoor)، حرم جامعي "أقرب للعالم الحقيقي".
- طبيعة البيانات: صور RGB.
- عدد الكاميرات: 8 كاميرات مراقبة غير متداخلة.
- الحجم: حوالي 36,411 صورة لـ 1,812 هوية (702 للتدريب، والباقي للاختبار) (CVF Open Access).

السمات المميِّزة:

- مشاهد أكثر ازدحامًا، مع تغيّر واضح في الإضاءة وزوايا التصوير.
- مستوى أعلى من التعقيد مقارنة بـ 1501-Market، ما يجعلها معيارًا أقوى لاختبار قدرة النماذج على التعميم.

أفضل حالات الاستخدام:

- تقييم الاستقرارية عبر تغيّر الكاميرا والزوايا .(Cross-View Robustness).
- اختبار أداء النماذج في بيئات أكثر واقعية، خاصة عند الانتقال من تجارب Market-1501 إلى بيئة أكثر تعقيدًا.

3.3.4. مجموعة بيانات DukeMTMC-VideoReID [11]

الجهة المطوِّرة:

تم تقديم DukeMTMC-VideoReID في ورقة Exploit the Unknown Gradually: One-Shot Video-Based Person Re-Identification by Stepwise Learning بواسطة Wu وزملائه في مؤتمر CVPR 2018. (CVF Open Access)

نظرة عامة:

هي نسخة تعتمد على الفيديو من DukeMTMC، حيث تُجمَّع إطارات الشخص الواحد عبر الزمن في مسارات (Tracklets)، بهدف تقييم نماذج ReID المعتمدة على التسلسل الزمني للحركة.

الخصائص الرئيسة:

- النوع: فيديو (Video-based ReID) قائم على المسارات.
- البيئة: خارجي (Outdoor)، حرم جامعة Duke.
- طبيعة البيانات: مسارات فيديو RGB (تتابعات من الإطارات).
- عدد الكاميرات: 8.
- الحجم: حوالي 4,832 مسار فيديو لـ 702 هوية، بمتوسط يقارب 168 إطارًا لكل مسار(GitHub) .

السمات المميِّزة:

- استمرارية زمنية واضحة لكل مسار تحت كل كاميرا.
- إمكانية استغلال المعلومات الزمنية (Temporal Cues) مثل نمط المشي والتغيّر التدريجي في الوضعيات.

أفضل حالات الاستخدام:

- تدريب وتقييم نماذج ReID المعتمدة على الفيديو مثل (3D CNN، وAttention Zمني، وTemporal Pooling).

- دراسة تأثير طول المسار وتغيّر الإطار عبر الزمن على جودة التمثيل المميِّز للشخص.

3.3.5. مجموعة بيانات (MARS: Motion Analysis and Re-identification Set) [12]

الجهة المطوِّرة:

قدّمها Liang Zheng وزملاؤه في ورقة MARS: A Video Benchmark for Large-Scale Person Re-identification في مؤتمر ECCV 2016، باعتبارها امتدادًا بالفيديو لمجموعة Market-1501.

(researchportalplus.anu.edu.au)

نظرة عامة:

تُعتبر MARS من أكبر مجموعات البيانات الخاصة بإعادة التعرّف من خلال الفيديو، حيث تم استخدام كاشف DPM ومتتبّع GMMCP لبناء مسارات الأشخاص تلقائيًا في بيئة حرم جامعي خارجي.

الخصائص الرئيسة:

- النوع: فيديو (Video-based ReID).
- البيئة: خارجي (Outdoor)، بيئة شبيهة بـ Market-1501.
- طبيعة البيانات: مسارات فيديو RGB.
- عدد الكاميرات: **6**.
- الحجم: حوالي 20,000 مسار فيديو لـ 1,261 هوية وأكثر من 1.1 مليون إطار.

(researchportalplus.anu.edu.au)

السمات المميّزة:

- وجود ضوضاء في الكشف والتتبع (أخطاء في Bounding Boxes، انسدادات، وتتبع غير مثالي).
- تشابه كبير مع الظروف العملية لأنظمة المراقبة الحقيقية.

أفضل حالات الاستخدام:

- تدريب نماذج فيديو قوية تتحمّل الضوضاء في الكشف والتتبع.
- تقييم النماذج التي تستفيد من كلٍّ من المعلومة المكانية (Appearance) والزمنية (Motion) في آن واحد.

3.3.6. مجموعة بيانات ICFG-PEDES (Identity-Centric and Fine-Grained PEDES) [13]

الجهة المطوِّرة:

قدمها Zefeng Ding وزملاؤه في ورقة Semantically Self-Aligned Network for Text-to-Image Part-aware Person Re-identification (SSAN) عام 2021، مع تقديم قاعدة البيانات الجديدة ICFG-PEDES كقاعدة نص–صورة مخصّصة لـ ReID. (arXiv)

نظرة عامة:

تم تصميم ICFG-PEDES لمهمة Text-to-Image ReID، حيث يتم استرجاع صورة الشخص من قاعدة صور كبيرة اعتمادًا على وصف نصي طبيعي. تستند الصور إلى مجموعة 17MSMT لكنها زوِّدت بوصف نصي دقيق لكل صورة. (ScienceDirect)

الخصائص الرئيسة:

- النوع: نص + صورة (Caption-based / Multimodal ReID).
- البيئة: خارجي (Outdoor)، مشاهد مراقبة متنوعة من MSMT17.
- طبيعة البيانات: صور RGB + أوصاف نصية باللغة الإنجليزية.

- الحجم: حوالي 54,522 زوج صورة–نص لـ 4,102 هوية، مقسّمة إلى مجموعة تدريب (34,674 زوجًا) واختبار (19,848 زوجًا). (ScienceDirect)

السمات المميّزة:

- الأوصاف النصية مركّزة على الهوية ودقيقة (Identity-Centric & Fine-Grained).

(opendatalab.com)

- طول الأوصاف أكبر وأكثر تفصيلاً من بعض قواعد البيانات الأخرى مثل (CUHK-PEDES)، ما يجعلها مناسبة لتعلّم تمثيلات لغوية–بصرية غنية.

أفضل حالات الاستخدام:

- تدريب نماذج Text-to-Image Person Search.
- بناء نماذج متعددة الوسائط (Vision–Language) تستخدم الوصف النصي لتوجيه استرجاع الشخص، مثل RaSa، SSAN، وغيرها. (arXiv)

3.3.7. مجموعة بيانات CUHK03 [14]

الجهة المطوِّرة الأصلية:

Wei Li وزملاؤه، في ورقة DeepReID: Deep Filter Pairing Neural Network for Person Re-identification (مؤتمر CVPR 2014) ، حيث قُدِّمت مجموعة بيانات CUHK03 كواحدة من أوائل المجموعات واسعة النطاق لإعادة التعرّف من صورة واحدة مع لقطات يدويّة وتلقائية. (cv-foundation.org)

نظرة عامة:

- النوع: صورة ثابتة (Image-based ReID).
- البيئة: حرم جامعي (مختلط داخلي/خارجي)، ضمن جامعة هونغ كونغ الصينية.
- طبيعة البيانات: صور RGB لأشخاص ملتقطة من أزواج كاميرات.

- عدد الكاميرات: 6 كاميرات
- الحجم: حوالي 14,097 صورة لـ 1,467 هوية، مع كل هوية تظهر في زوج من الكاميرات.

السمات المميّزة:

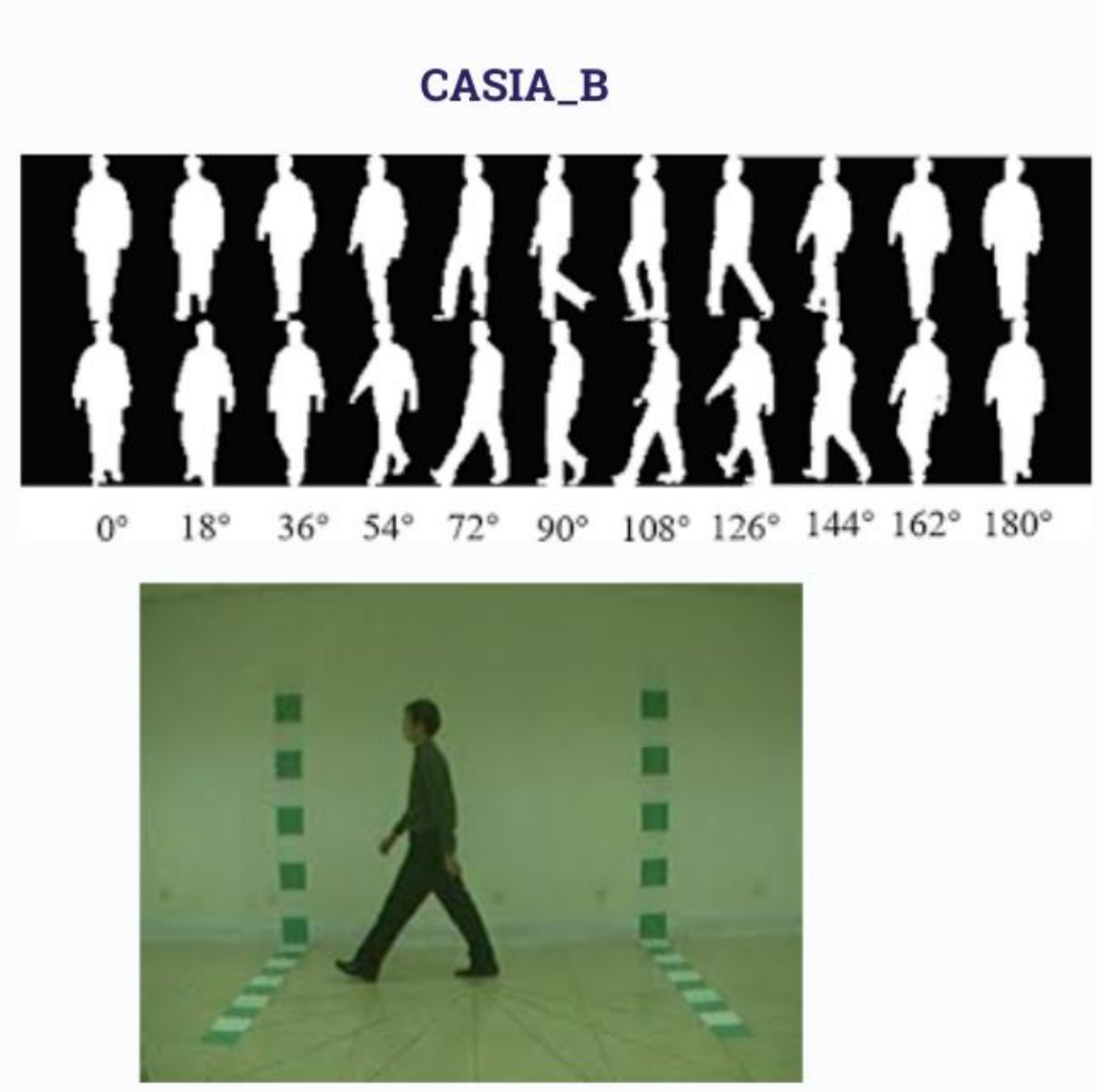
- توفر نوعين من مربعات الإحاطة:
  1. يدوية (Labeled) – مقصوصة يدويًا بدقة عالية.
  2. تلقائية (Detected) – ناتجة عن كاشف DPM لمحاكاة النظام الحقيقي. (cv-foundation.org)
- تُعد من أوائل المجموعات التي ركّزت على تحدّي تشابه الملابس بين الطلبة داخل الحرم الجامعي ووجود تغيّر في الزوايا والمسافة.
- نسخة TIP-CB على Kaggle تُوفّر تقسيمًا جاهزًا للتدريب/الاختبار، ما يسهّل الدمج السريع في تجاربك. (Kaggle)

3.3.8. مجموعات بيانات المشية لإعادة التعرف

مجموعة بيانات CASIA-B (Gait Dataset B) [15]

- تُعد CASIA-B واحدة من أقدم وأكثر مجموعات بيانات المشية تأثيرًا في تاريخ المجال، وغالبًا ما يُشار إليها بوصفها المرجع التاريخي والخط الأساس (Historical Baseline) الذي تأسست عليه معظم أبحاث التعرّف على المشية اللاحقة. طُوِّرت هذه المجموعة في معهد علوم الأتمتة بالأكاديمية الصينية للعلوم (CASIA)، وقد لعبت دورًا محوريًا في توحيد بروتوكولات التقييم ووضع أسس المقارنة العادلة بين النماذج المختلفة.
- تضم CASIA-B تسلسلات مشية لعدد يقارب 124 شخصًا، مصوّرة من 11 زاوية رؤية مختلفة تتراوح من 0° إلى 180°، مع توفير ثلاثة سيناريوهات مشي مميّزة: المشي الطبيعي، المشي أثناء حمل حقيبة، والمشي أثناء ارتداء معطف. هذا التنويع المنهجي في ظروف المشي جعل المجموعة مناسبة لدراسة تأثير العوامل الخارجية على نمط المشية، مثل تغيّر الشكل الخارجي أو إعاقة حركة الأطراف، وهو ما لا يتوافر بنفس الوضوح في كثير من المجموعات الأحدث.
- تُستخدم CASIA-B على نطاق واسع في تدريب وتقييم نماذج المشية المعتمدة على الظلال مثل GaitSet وGaitGL، كما تُعد مرجعًا إلزاميًا للإبلاغ عن النتائج الأساسية في معظم الأوراق العلمية، حتى تلك التي تستهدف تطبيقات أكثر واقعية. ومع ذلك، فإن بيئتها المتحكَّم بها نسبيًا، وعدد الهويات المحدود مقارنة بالمجموعات الحديثة،

يفرضان قيودًا على قدرتها على تمثيل سيناريوهات المراقبة الحقيقية. ورغم هذه القيود، ما تزال CASIA-B تحتفظ بمكانتها العلمية، ليس بوصفها اختبارًا للجاهزية الواقعية، بل باعتبارها نقطة الانطلاق القياسية التي تُمكّن من قياس التقدّم المنهجي في نماذج التعرّف على المشية عبر الزمن.



الشكل 11 مجموعة بيانات CASIA-B الزوايا وطريقة التقاط الصور

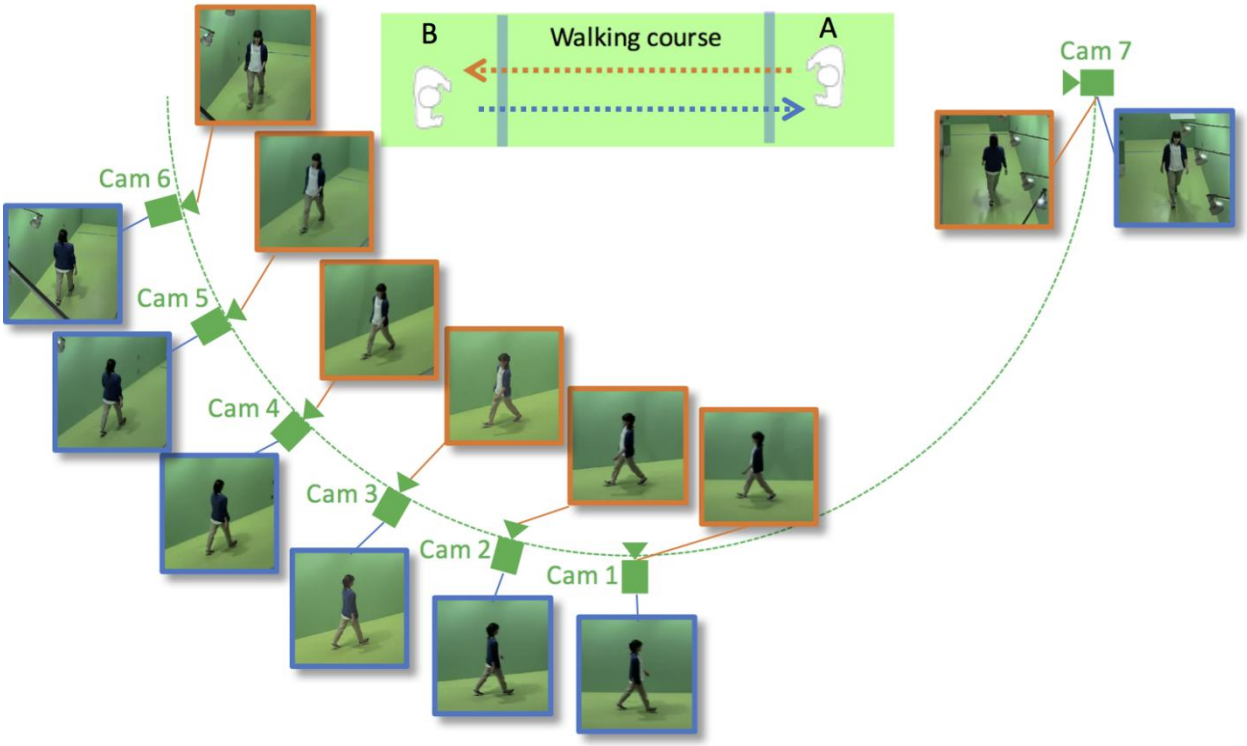
يذكر أن النسخة ذات الصور الخام RGB غير متوفرة للتحميل وما هو متوفر هي الصور الظلية المستخرجة منها دون الأصل

مجموعة بيانات **OU–MVLP** [16]

تُعد OU-MVLP (Osaka University – Multi-View Large Population) واحدة من أهم وأكبر مجموعات البيانات المرجعية عالميًا في مجال التعرّف على المشية المعتمد على الظلال، وقد أصبحت معيارًا فعليًا (de-facto benchmark) لأي بحث يتناول **مشكلة اختلاف زوايا الرؤية** (View Variance). تتميّز هذه المجموعة بحجمها الضخم الذي يضم أكثر من 10,000 هوية بشرية، وهو ما يجعلها من أوائل مجموعات بيانات المشية التي انتقلت بالمجال من نطاق التجارب المحدودة إلى النطاق الإحصائي واسع التمثيل.

تحتوي OU-MVLP على تسلسلات مشية ملتقطة من 14 زاوية تصوير مختلفة موزعة بانتظام حول الشخص (من 0° إلى 180°)، مع توفّر **تمثيلات ظلية** (Silhouettes) عالية الاتساق لكل زاوية، الأمر الذي يسمح بدراسة المشية بوصفها خاصية بيوميكانيكية مستقلة عن زاوية الكاميرا. هذا التصميم يجعل المجموعة مناسبة بشكل خاص لتدريب وتقييم النماذج التي تهدف إلى تعلّم تمثيلات مشية غير متحيّزة للرؤية، مثل نماذج التحويل بين الزوايا، والنماذج المعتمدة على الانتباه الزمني، ومُرمِّزات المشية المبنية على بنى Transformers.

وعلى خلاف مجموعات البيانات الأقدم مثل CASIA-B، لا تركز OU-MVLP على تنويع ظروف الحمل أو الملابس، بل تضع تركيزها الأساسي على تفكيك أثر زاوية الرؤية بوصفه أحد أكثر التحديات تعقيدًا في أنظمة المشية الواقعية. لذلك، يُنظر إليها في الأدبيات على أنها الاختبار الحاسم لأي ادعاء بعدم التأثر بزاوية التصوير، ويُتوقّع من أي نموذج يقدّم هذا الادعاء أن يُقيَّم على OU-MVLP. ورغم أن بيئتها ما تزال شبه مُتحكَّم بها، فإن حجمها وتنوّع زواياها جعلاها حجر الأساس لتطوير نماذج المشية الحديثة، ومكوّنًا شبه إلزامي في بروتوكولات التقييم المعتمدة في أبحاث التعرّف على المشية وإعادة التعرّف القائمة عليها.



الشكل 12مجموعة بيانات OU-MVLP الزوايا وطريقة التقاط الصور

مجموعة بيانات GREW [17]

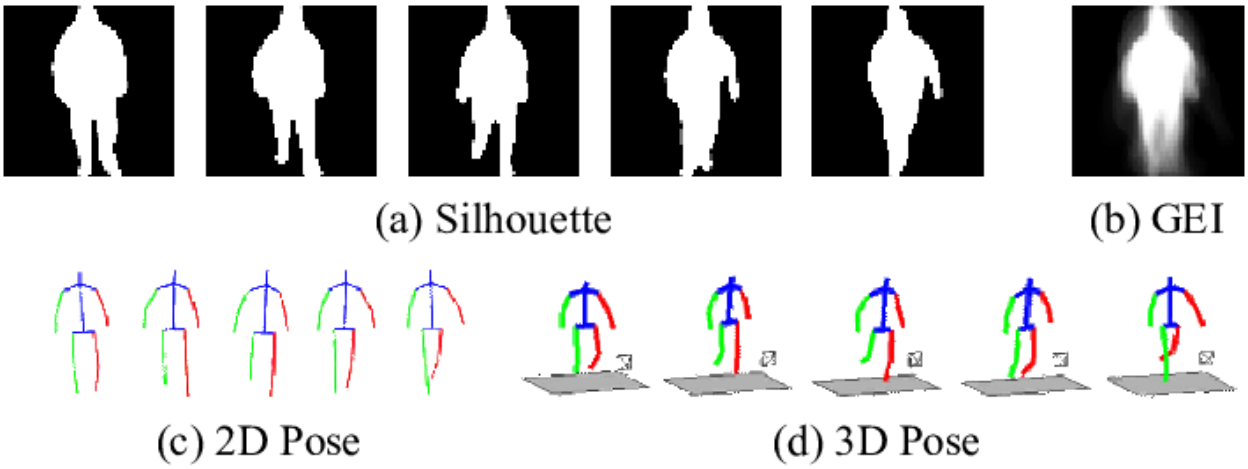
تُعد GREW (Gait Recognition in the Wild) واحدة من أهم مجموعات البيانات الحديثة في مجال التعرّف على المشية، وتمثّل نقطة التحوّل الأساسية من البيئات المتحكَّم بها إلى بيئات المراقبة الواقعية. طُوِّرت GREW بهدف سدّ الفجوة بين مجموعات البيانات الأكاديمية التقليدية وتحديات النشر الفعلي، إذ تعتمد على لقطات مراقبة حقيقية مأخوذة من أنظمة CCTV في بيئات مفتوحة وغير مُنضبطة.

تتميّز GREW بحجمها الكبير الذي يضم عشرات الآلاف من الهويات ومئات الآلاف من تسلسلات المشي، مع تنوّع واسع في زوايا التصوير، والمسافات، وجودة الصورة، والازدحام، والانسدادات الجزئية (Occlusions). وعلى عكس CASIA-B وOU-MVLP، لا توفّر GREW زوايا تصوير مُعرّفة مسبقًا أو ظروفًا مُتحكَّم بها، بل تترك هذه العوامل لتتقلب طبيعيًا كما هو الحال في أنظمة المراقبة الحقيقية. ونتيجةً لذلك، تُجبر النماذج المدرَّبة أو المقيَّمة عليها على التعامل مع الضجيج البصري، وعدم استقرار المشية، وتفاوت أطوال المقاطع الزمنية.

تُستخدم GREW على نطاق واسع في اختبار المتانة (Robustness)، وفي التدريب المسبق واسع النطاق لنماذج المشية الحديثة، كما تُعد معيارًا أساسيًا لأي عمل بحثي يدّعي الجاهزية للتطبيق العملي أو الصناعي. وغالبًا ما تُستخرج منها تمثيلات الظلال، بينما تُستخدم الوضعية (Pose) أو السمات الحركية كمصادر إشراف مساعدة ضمن نماذج هجينة متعددة الأنماط. وبسبب طبيعتها الواقعية، لا تُعد GREW مجرد مجموعة تقييم إضافية، بل تُنظر إليها في الأدبيات بوصفها الاختبار الحاسم لقدرة نماذج المشية على الانتقال من المختبر إلى الواقع، وهو ما منحها مكانة مرجعية راسخة في الأبحاث الحديثة الخاصة بإعادة التعرّف المعتمدة على المشية.

إن الاختلاف الأساسي بين مجموعة البيانات GREW وبين مجموعات البيانات الفيديوية هي أنها تضمن المشيات دوماً، لا توجد فيها سلاسل أطر لأشخاص متسمرين في اماكنهم، كما أنها تضمن أن كافة المرورات tracklets تحوي كافة أطوار المشية الواحدة مرة واحدة على الاقل لذلك تصنف على انها مجموعة بيانات مشية Gait based.

قام العديد من الباحثين باشتقاق مجموعات بيانات منها بعد استخراج الصور الظلية والوضعيات ثنائية وثلاثية الابعاد



الشكل 13 المعطيات الظلية والوضعيات في الفضائين ثنائي الابعاد وثلاثس الابعاد المشتقى من صور محموعة البيانات GREW

3.4. المشية والخصائص البيومترية المرتبطة بها

3.4.1. مقدمة

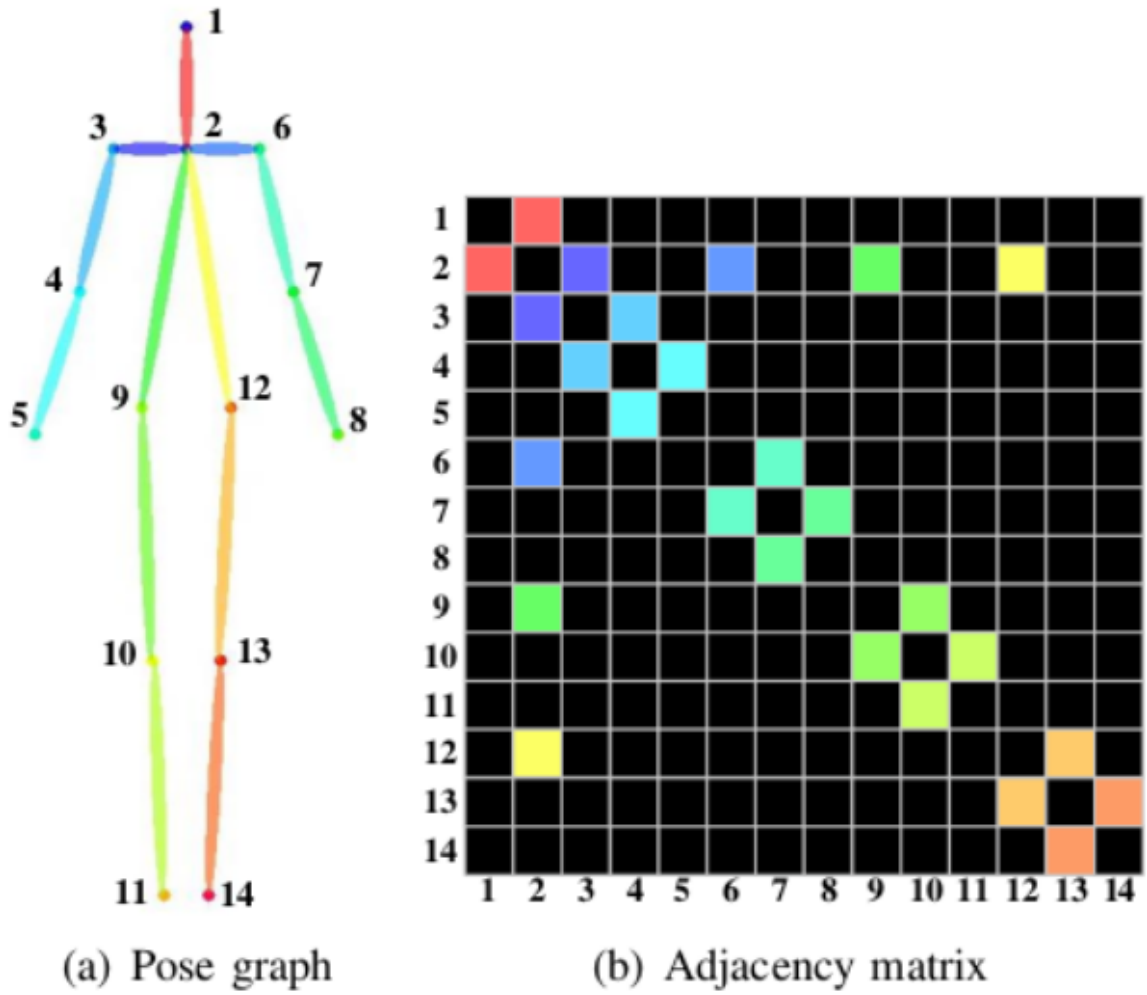
سنناقش النماذج التي اعتمد على وضعية الجسم pose فقط ثم تلك المعتمدة على الصور الظلية ثم المقاربات التي حاولت الدمج فيما بينها.

من الضروري أن نذكر أن هذه النماذج تعتمد على مجموعات البيانات المعيارية من صنف Gait based.

تُعدّ المشية بطبيعتها معطىً ذا طبيعة زمنية، حيث تكون وحدة الدخل الأساسية في هذا السياق هي تسلسل من الأطر المترابطة زمنيًا (Tracklet) يمثّل مرور شخص واحد أمام كاميرا معيّنة، ويغطي دورة مشي واحدة على الأقل. وبناءً على ذلك، لا يمكن توصيف المشية توصيفًا كافيًا اعتمادًا على إطار منفرد، بل يتطلّب الأمر تحليل التغيّر الحركي عبر الزمن.

وفقًا لنوع التمثيل المعتمد، يكون دخل النموذج إما تسلسلًا من الصور الظلية في حال كانت المقاربة موجّهة لالتقاط نمط تغيّر شكل أو محيط الجسم أثناء المشي، أو تسلسلًا من البنى الممثِّلة لوضعية الجسم. ومن أشهر هذه البنى تمثيل COCO-17، الذي يعبّر عن الجسم البشري باستخدام مجموعة من المفاصل (Joints أو Nodes) وروابط فيما بينها تمثّل العظام (Edges). وتتميّز هذه البنية بأن العلاقات المكانية بين أجزاء الجسم ثابتة بنيويًا، في حين تتغيّر المواضع النسبية للعقد وأطوال الوصلات مع الزمن نتيجة الحركة.

وبسبب هذه الخصائص، يُنظر إلى تمثيل الوضعيات على أنه بنية بيانية (Graph Structure)، وغالبًا ما تُعبَّر عنها باستخدام مصفوفات تجاور (Adjacency Matrices) تصف العلاقات بين المفاصل. وعند التعامل مع تسلسل زمني من هذه البنى، فإننا نكون أمام بيان متغيّر زمانيًا ومكانيًا، تتبدّل فيه سمات العقد والوصلات من إطار إلى آخر ضمن التسلسل.



الشكل 14 تمثيل بيان وضعية جسم الانسان في مصفوفة تجاور

3.4.2. التغيّر في وضعية الجسم (Pose) خلال المشي كخاصية مميّزة

تعتمد النماذج المعتمدة على الوضعية (Pose-based Gait Models) على استخراج وضعية الجسم لكل إطار باستخدام نماذج تقدير الوضعية البشرية الحديثة، مثل HRNet أو YOLO-Pose، ثم تمرير البنية البيانية الناتجة كدخل إلى نموذج التعلّم. ويتمثّل البعد الزمني في هذه النماذج في الفروقات المتعاقبة في مواقع المفاصل وأطوال الوصلات بين بيان وآخر، أي في التغيّر المستمر للتشكيل الهندسي للجسم أثناء المشي.

المشية كوصف حركي قائم على المفاصل (Pose-based Gait as Joint Motion Descriptors) [18]

مثّل الانتقال من النماذج المعتمدة على الصور الظلية إلى النماذج المعتمدة على وضعية الجسم تحوّلًا مفاهيميًا مهمًا في مجال التعرّف على المشية، حيث انطلقت هذه المقاربات من مسلّمة أساسية مفادها أن جوهر المشية يكمن في الحركة نفسها، لا في المظهر الخارجي للجسم. وقد جاء هذا التحوّل مدفوعًا بمحدودية النماذج الظلية في التعامل مع تغيّرات الملابس، وحمل الأغراض، والتداخل مع الخلفية، فضلًا عن اعتمادها الضمني على سمات مظهرية لا تعبّر بالضرورة عن نمط المشي.

في هذا السياق، قدّم Liao وآخرون (2020) نموذج PoseGait بوصفه أحد أوائل الأعمال التي اعتمدت بشكل صريح على وضعية الجسم (Pose) كمصدر أساسي لتمثيل المشية، دون الاستناد إلى الصور الظلية أو السمات المظهرية. تنطلق هذه الورقة من فرضية واضحة مفادها أن التغيّر الزمني في مواقع المفاصل البشرية يحمل معلومات كافية لتمييز الأفراد، وأنه يمكن استغلال هذه المعلومات لبناء تمثيل مقاوم لتغيّرات المظهر.

يعتمد PoseGait على استخراج الوضعية البشرية لكل إطار ضمن تسلسل المشي باستخدام نماذج تقدير الوضعية، ثم تمثيل المشية على شكل سلسلة زمنية من المفاصل. ويتم توصيف كل مفصل من خلال إحداثياته المكانية، مع اشتقاق سمات حركية إضافية مثل الإزاحات، والسرعات، والزوايا النسبية بين المفاصل. ويُنظر إلى المشية في هذا النموذج على أنها تجميع زمني لهذه السمات المفصلية، حيث يُمثّل البعد الزمني من خلال التغيّر العددي في مواقع المفاصل عبر الإطارات المتتالية.

تعكس هذه المقاربة تصورًا مبكرًا للمشية بوصفها سلوكًا حركيًا يمكن اختزاله إلى متجهات سمات، حيث تُعامل المفاصل كوحدات مستقلة نسبيًا، وتُفترض العلاقات التشريحية بينها على نحو ضمني، استنادًا إلى المعرفة المسبقة بجسم الإنسان. وعلى الرغم من إدخال مفهوم الزمن، إلا أن هذا الزمن يُتعامل معه كمتغيّر حسابي (differences over time)، لا كبنية متعلّمة بحد ذاتها.

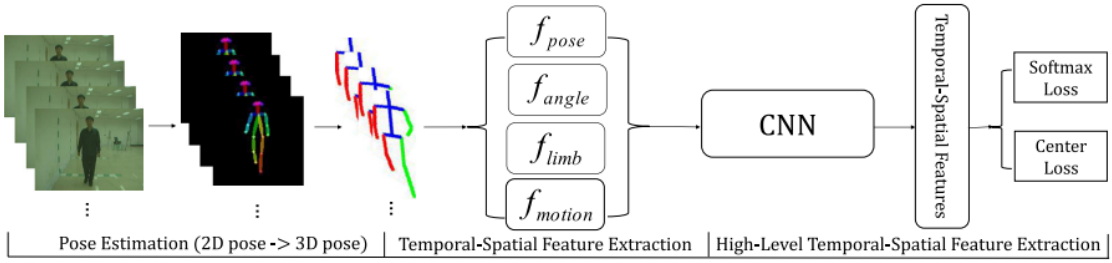
تكمن أهمية PoseGait في كونه قد أسّس لعدة مسلّمات ستؤثّر لاحقًا في تطوّر المجال، أهمها:

1. إمكانية فصل المشية عن المظهر الخارجي باستخدام الوضعية فقط.
2. اعتبار التسلسل الزمني عنصرًا لا غنى عنه في توصيف المشية.
3. الاعتماد على معرفة تشريحية مسبقة لتوجيه عملية استخراج السمات.

إلا أن هذه المسلّمات نفسها كشفت عن حدود هذا التصوّر. فغياب تمثيل صريح للعلاقات المكانية بين المفاصل، والاكتفاء بتوصيفها ضمنيًا عبر السمات، أدّى إلى فقدان القدرة على نمذجة التفاعل البنيوي بين أجزاء الجسم. كما أن معالجة الزمن بوصفه محورًا حسابيًا، لا كبنية ديناميكية، حدّت من قدرة النموذج على التقاط الأنماط الحركية المعقّدة التي تميّز المشية البشرية.

وبذلك، يمكن النظر إلى PoseGait على أنه مرحلة انتقالية في تطوّر النماذج المعتمدة على الوضعية: مرحلة نجحت في كسر الارتباط بين المشية والمظهر، لكنها لم تصل بعد إلى تمثيل المشية بوصفها بنية حركية مترابطة. وقد مهّدت هذه المحدوديات مباشرة لظهور أعمال لاحقة سعت إلى تمثيل الهيكل العظمي كبنية بيانية صريحة، وإعادة صياغة المشية على أنها تسلسل من العلاقات المكانية—الزمانية المتعلَّمة.

استخدم هذا النموذج بنى تقليدية (CNN) وهي من النماذج التي يشيع استخدامها في مسائل الرؤية الحاسوبية للنمذجة إضافة إلى نمذجة التغيرات الزمانية والمكانية



الشكل 15 بنية نموذج POSEGAIT

لم يستطع هذا النموذج التقاط الترابط البنيوي الذي يتمتع به جسم الانسان عل عكس النموذج PoseGait

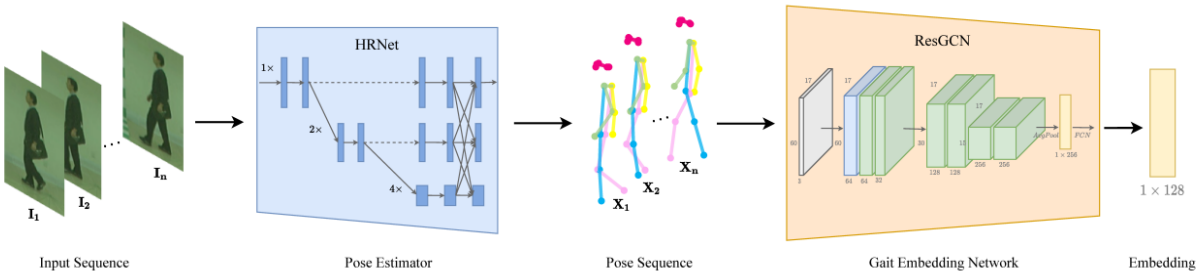
جاءت ورقة GaitGraph [19] التي قدّمها Teepe وآخرون عام 2021 بوصفها استجابة مباشرة للقيود البنيوية والمفاهيمية التي كشفت عنها النماذج المعتمدة على الوضعية من الجيل الأول، وعلى رأسها PoseGait. وتنطلق هذه الورقة من مسلّمة أعمق مفادها أن المشية لا يمكن اختزالها إلى مجرد تغيّر عددي في مواقع المفاصل، بل هي ناتج تفاعل بنيوي منسّق بين أجزاء الجسم عبر الزمن، وأن هذا التفاعل يمكن تمثيله على نحو طبيعي ودقيق باستخدام بنى بيانية (Graphs). بناءً على ذلك، يعيد GaitGraph صياغة مشكلة التعرّف على المشية باعتبارها مسألة تعلّم على تسلسل من الرسوم البيانية الهيكلية المتغيّرة زمانيًا، حيث يُمثَّل كل إطار من تسلسل المشي على شكل بيان ثابت الطوبولوجيا يعكس البنية التشريحية لجسم الإنسان، بينما تعبّر السمات المرتبطة بالعقد عن مواقع المفاصل وثقتها، ويتجسّد البعد الزمني في تعاقب هذه الرسوم عبر الزمن.

من الناحية المعمارية، يعتمد النموذج على شبكات التفاف بيانية مكانية—زمانية (Spatio-temporal GCNs)، حيث تُستخدم عمليات التفاف بياني مكاني لالتقاط العلاقات التشريحية بين المفاصل ضمن كل إطار، تليها عمليات التفاف زمنية لنمذجة تطوّر هذه العلاقات عبر الإطارات المتتالية. يمثّل هذا الفصل الواضح بين البعد المكاني والبعد الزمني تحوّلًا مفاهيميًا مهمًا مقارنةً بالأعمال السابقة، إذ لم يعد الزمن مجرّد محور حسابي للفروقات، بل أصبح بُعدًا متعلَّمًا داخل النموذج. كما يتيح

استخدام البنية البيانية استغلال التفاعل التعاوني بين المفاصل، مثل تناغم الساقين وحركة الذراعين المتزامنة مع الجذع، وهو ما كان غائبًا أو ضمنيًا في النماذج السابقة.

على المستوى التجريبي، أظهرت نتائج GaitGraph على مجموعة بيانات CASIA-B تحسّنًا كبيرًا مقارنةً بنموذج PoseGait، حيث حقّق النموذج دقة Rank-1 بلغت 87.7% في حالة المشي الطبيعي (NM)، و74.8% عند حمل الأغراض (BG)، و66.3% عند ارتداء المعاطف (CL)، أي بزيادة تجاوزت 19 نقطة مئوية في حالة NM وأكثر من 30 نقطة في حالتي BG وCL مقارنةً بـ PoseGait. وتُعد هذه القفزة العددية دلالة واضحة على أن التحسّن لم يكن ناتجًا عن تعقيد معماري فحسب، بل عن تغيّر جذري في تمثيل المشية ذاتها. ويعزّز هذا الاستنتاج ما قدّمته الورقة من تجارب إزالة (Ablation Studies)، حيث أدّى خلط الترتيب الزمني للإطارات إلى تدهور حاد في الأداء، ما يثبت أن النموذج يعتمد فعليًا على الديناميكا الزمنية للمشية، لا على إحصاءات وضعية ساكنة.

مع ذلك، ورغم هذا التقدّم، يحتفظ GaitGraph بعدد من الافتراضات التي ستصبح لاحقًا موضع نقد في الأعمال الأحدث. إذ تفترض البنية البيانية أن جميع المفاصل والعلاقات التشريحية متساوية الأهمية في توصيف المشية، كما تعتمد على طوبولوجيا ثابتة لا تتكيّف مع السياق الحركي أو مع اختلاف الأهمية الدلالية للأطراف المختلفة أثناء المشي. وبالرغم من هذه القيود، يُنظر إلى GaitGraph على أنه نقطة التحوّل المفصلية التي نقلت النماذج المعتمدة على الوضعية من توصيف حركي مفصلي إلى نمذجة بنيوية مكانية–زمانية متكاملة، ومهّد الطريق لظهور مقاربات أكثر نضجًا تعتمد على آليات الانتباه والعلاقات الدلالية، كما في النماذج المعتمدة على Graph Transformers في الأعمال اللاحقة.



الشكل 16بنية نموذج GAITGRAPH

اعتمد هذا النموذج عل بنية ResGCN [20] التي أتاحت كانت أكثر اتساقا فلسفة الأولى رفض النظر إلى المفاصل كوحدات مستقلة.

في المقاربات السابقة (مثل PoseGait)، كان التفكير السائد هو:

كل مفصل يتحرّك، ونقيس حركته عبر الزمن، ثم نُجمّع هذه القياسات.

أما ResGCN فتنطلق من مسلّمة مختلفة:

وهي كون المفصل لا معنى لحركته إلا ضمن علاقته بالمفاصل الأخرى.

بالتالي، التمثيل الصحيح للمشية لا يبدأ من الحركة، بل من البنية التي تتحرّك وليست العقد والوصلات مستقلة تماماً عن بعضها بسبب طبيعة الحركة ذاتها وإن كنا ننظر عادة إلى العقد والوصلات على أنها أكثر استقلالاً بتعريف البيان كبنية.

على الرغم من أن ResGCN—كما استُخدمت في نماذج مثل GaitGraph (2021) —مثّلت قفزة نوعية عبر تمثيل المشية كبنية مكانية–زمانية صريحة، إلا أن هذا النجاح كشف عن حدود فلسفية وبنيوية لا يمكن تجاوزها ضمن إطار GCN التقليدي.

أول هذه الحدود هو الطوبولوجيا الثابتة؛ إذ تفترض ResGCN أن العلاقات التشريحية بين المفاصل ثابتة وكافية لتوصيف المشية، بينما تُظهر المشية في الواقع أن العلاقات الأكثر دلالة قد تكون غير محلية (مثل تناغم الساقين عبر الجذع). ثانيًا، تفترض ResGCN تساوي الأهمية بين المفاصل، حيث تُجمَّع الإشارات عبر الجيران التشريحيين دون تمييز دلالي، وهو افتراض مناسب لإثبات أهمية البنية، لكنه غير كافٍ للتمييز الدقيق بين الأفراد. ثالثًا، يظلّ الزمن مُنمذجًا بإيقاع عام عبر التفاف زمني محلي، دون القدرة على التركيز الانتقائي على مقاطع حركية بعينها (Sub-patterns) ذات قيمة تمييزية أعلى. وأخيرًا، فإن طبيعة التجميع في GCN تجعل النموذج تجميعيًا (Averaging) أكثر منه تمييزيًا (Discriminative)، ما يحدّ من قدرته على إبراز العلاقات الحاسمة لهوية المشية.

وكما كان الاتجاه العام في خلال السنوات الماضية ومنذ ظهور المحولات transformers متجها إلى استكشاف قدراته في أغلب المسائل العتمدة على تعلم الآلة فقد جاءت Graph Transformers [21]استجابةً مباشرة لهذه الحدود المطروحة، ليس بوصفها تحسينًا معماريًا فحسب، بل كتحوّل في طريقة التفكير. الفكرة الجوهرية هي أن التمثيل الجيّد لا يساوي بين جميع العلاقات، بل يعيد وزنها وفق السياق. يتيح الانتباه (Attention) تعلّم اعتماديات بعيدة المدى بين المفاصل، وإعادة تشكيل البنية ديناميكيًا بدل الالتزام بجوار تشريحي ثابت. كما يسمح بالتمييز بين العلاقات ذات القيمة الدلالية وتلك الثانوية، ويجعل الزمن نفسه قابلًا لإعادة الوزن، لا مجرد محور ترتيب. بهذا، تنتقل النمذجة من «الهندسة» (من متصل بمن) إلى الدلالة (من يؤثّر على من ولماذا)، وهو انتقال ضروري حين يصبح هدف النموذج تمييز الهوية لا مجرد توصيف الحركة.

ظهرت في عام الورقة البحثية التي اقترحت نموذجاً جديداً متفقاً مع هذا التوجه وهي MoCos (Motif-guided Collaborative Skeleton Modeling, 2025) [22] التي انطلقت من مسلّمة أعمق من سابقاتها :

المشية ليست تفعيلًا لبنية تشريحية فحسب، بل تنظيمًا دلاليًا تعاونيًا بين مجموعات مفاصل عبر الزمن.

أي أن القيمة التمييزية لا تكمن في مفصل منفرد، بل في أنماط التعاون (Motifs) بين مفاصل محددة (مثل تناغم الساقين أو اقتران الساق مع الجذع)، وهي أنماط مستمدّة من علم المشي (Biomechanics).

وقدمت الاسهامات التالية:

- دمج Graph Transformers لإعادة وزن العلاقات ديناميكيًا بدل تجميعها بالتساوي.
- توجيه الانتباه بالموتيفات: الانتباه ليس حرًّا تمامًا، بل مُقَيَّد بدلالات حركية معروفة، ما يقلّل الضجيج ويزيد التفسيرية.
- نمذجة زمنية أدق: الانتقال من إيقاع عام إلى مقاطع حركية جزئية يُعاد وزنها حسب أهميتها.
- تحويل المشية إلى تمثيل دلالي: من Semantics ⟶ Pose ⟶ Graph ⟶ Attention، وهو نضج مفاهيمي واضح في تعريف المشكلة.

ولكن وبسبب اعتمادها على بنية المحولات فقد جاء هذا مرافقاً لتعقيد حسابي أعلى بالمقارنة مع النماذج السابقة. كما أنها ما زالت كسابقاتها متأثرة بجودة حساب الوضعية التي تتأثر بدورها بجودة الصور وعدم وجود حجب لأجزاء من الجسم من عوائق او من لباس معين كمعطف طويل او حقيبة يد.

فيما يلي جداول مقارنة بين أجيال النماذج التي قمنا باستعراضها في هذا الفصل

النموذج	سنة الإصدار	التوجه المفاهيمي
PoseGait	2020	المشية كسلسلة خصائص حركية للعقد
GaitGraph	2021	المشية كبنية زمانية مكانية
MoCos	2025	المشية كنمط تعاوني دلالي موجه بالانتباه

الجدول 5 أجيال النماذج المعتمدة على تغير الوضعيات ومقارباتها

النموذج	البنية	تمثيل الوضعية أو الهيكل
PoseGait	CNN	عقد مستقلة مسماة ذات احداثيات
GaitGraph	ResGCN	بيان
MoCos	محول بياني	بيان ديناميكي دلالي + آلية انتباه

الجدول6 أجيال النماذج المعتمدة على تغير الوضعيات البتية وتمثيل المعطيات

جميع النماذج اعتمدت على مجموعة البيانات Casia-B

النتائج Rank-1:

النموذج	ملابس عادية NM	مع حقيبة يد BG	مع معطف طويل CL
PoseGait	68.7	44.5	36.0
GaitGraph	87.7	74.8	66.3
MoCos	87.9	73.6	72.1

الجدول 7 أجيال النماذج المعتمدة على تغير الوضعيات مقارنة جودة النتائج

النموذج	**القيمة المضافة الأساسية**	**التحديات غير المحلولة**
PoseGait	فصل المشية عن المظهر؛ إدخال الـ Tracklet الزمني	لا بنية صريحة؛ زمن غير متعلّم
GaitGraph	نمذجة بنيوية مكانية—زمانية؛ قفزة أداء كبيرة	طوبولوجيا ثابتة؛ تساوي أهمية المفاصل
MoCos	دلالة حركية + انتباه موجَّه بالموتيفات	تعقيد حسابي؛ حساسية لجودة الوضعية

الجدول 8 أجيال النماذج المعتمدة على تغير الوضعيات والقيمة المضافة والتحديات

3.4.3. التغير في شكل الجسم (الصورة الظلية) كخاصية مميزة في مسألة إعادة التعرف

تُعدّ الصور الظلية (Silhouettes) أحد أقدم وأكثر التمثيلات استخدامًا في أنظمة التعرّف على المشية، إذ تهدف إلى عزل نمط الحركة الهندسية للجسم عن العوامل المربكة مثل اللون، الإضاءة، والخلفية. ويقوم هذا التمثيل على تحويل كل إطار فيديو إلى صورة ثنائية (أبيض/أسود) تُبرز حدود الجسم أثناء المشي.

تكمن الفلسفة الأساسية لهذا الاتجاه في أن المشية خاصية حركية بنيوية، وأن الشكل الديناميكي للجسم عبر الزمن يحمل معلومات تمييزية كافية حتى في غياب المظهر الخارجي.

قد يتماثل تمثيل وضعيتان لشخصين مختلفين لكن الصورة الظلية تظهر خصائص أخرى كثيرة مميزة مثل توزع كتلة الجسم هنا يمكننا التمييز بين شخصين بدين وآخر نحيل لهما الطول نفسه وهي خاصية ذات أهمية في إعادة التعرف دون شك.

غير أن هذا التبسيط، رغم فوائده، يفرض قيودًا بنيوية صارمة على النماذج التي تعتمد عليه، وهو ما يظهر بوضوح عند الانتقال من البيئات المخبرية إلى البيئات الواقعية.

ترافقت بداية الاهتمام بالتعلم العميق مع النماذج المعتمدة على الصور الظلية بتطبيق CNN قبل أن تبدأ النماذج المتناولة لهذه المسألة بطرح حلول أكثر ابتكاراً وتعقيداً

نموذج GaitSet: فلسفة "المشية كمجموعة" وحدودها [23]

قدّم نموذج GaitSet (2019) تحوّلًا مفاهيميًا بارزًا في نماذج التعرّف على المشية المعتمدة على الصور الظلية، إذ انطلق من فرضية أساسية مفادها أن المشية حركة دورية، وأن المظهر الشكلي لكل صورة ظلية يحمل ضمنيًا دلالة موقعها الزمني داخل دورة المشي، مما يجعل الترتيب الزمني الصريح للأطر غير ضروري فكانأن تعامل هذا النموذج ــ كما يقترح اسمه ــ مع أطر عنصر المرور كمجموعة غير مرتبة وأن الشعاع الممثل لهذا المرور هو محصلة لهذه الأطر من الصور الظلية بغض النظر عن ترتيبها، ويعتمد بنية تتكوّن من ثلاث مراحل رئيسية:

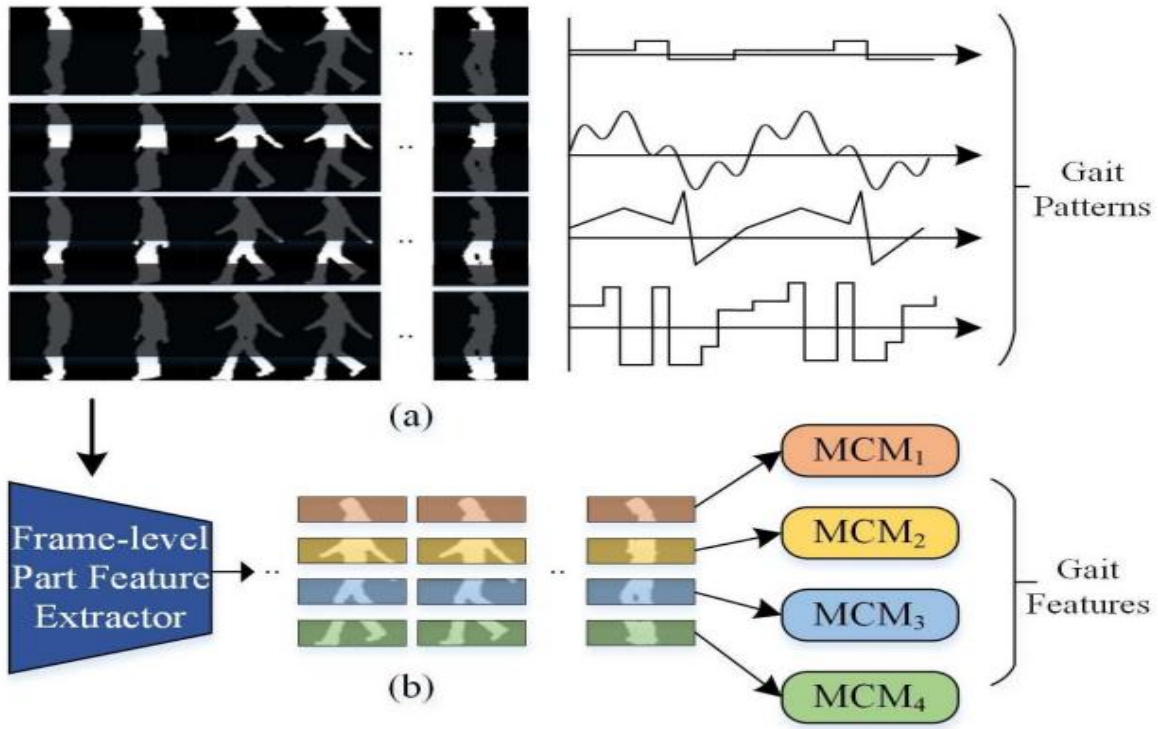
استخراج سمات إطار–بإطار باستخدام شبكة CNN مشتركة الأوزان، ثم تجميع هذه السمات عبر آلية Set Pooling المتوافقة مع خاصية عدم التأثر بترتيب العناصر (Permutation Invariance)، وأخيرًا إسقاط التمثيل الناتج إلى فضاء تمييزي باستخدام Horizontal Pyramid Mapping (HPM) لتعزيز السمات المحلية والعالمية في آن واحد. تعلّم النموذج تمّ اعتمادًا على صور ظلية مُحاذاة بدقة من مجموعات بيانات مقيدة واسعة الانتشار مثل CASIA-B و OU-MVLP، حيث حقق أداءً رائدًا وقت نشره، مسجلًا دقة Rank-1 بلغت 95.0% على CASIA-B و 87.1% على OU-MVLP في ظروف المشي الطبيعية، إضافةً إلى متانة ملحوظة في حالات حمل الأغراض وارتداء المعاطف، كما أظهر قدرة جيدة على العمل حتى

مع عدد محدود من الإطارات (نحو 7 صور فقط). تكمن القيمة المضافة الأساسية لـ GaitSet في بساطته البنيوية ومرونته العملية، إذ ألغى القيود الصارمة على طول التسلسل، وسمح بدمج صور من مقاطع مختلفة وزوايا متعددة ضمن تمثيل واحد، مما جعله نموذجًا مرجعيًا ومصدر إلهام لسلسلة واسعة من الأعمال اللاحقة. غير أن هذه الفلسفة ذاتها تمثل في الوقت نفسه موطن الضعف الجوهري للنموذج، إذ إن إهمال العلاقات الزمنية الصريحة يحرم الشبكة من نمذجة الديناميكيات الدقيقة للحركة، ويجعلها تميل إلى تعلّم أنماط شكلية شبه ساكنة بدل الخصائص الحركية الخالصة، وهو ما يفسّر تراجع أدائها بشكل ملحوظ عند الانتقال إلى مجموعات بيانات واقعية معقّدة (in-the-wild) حيث يصبح الترتيب الزمني، والتفاعل الحركي بين الإطارات، عاملين حاسمين لا يمكن تجاهلهما.

نموذج GaitPart (2020): من "المشية كمجموعة" إلى "المشية كأجزاء متحركة" [24]

الفلسفة والمنطلق النظري

ينطلق نموذج GaitPart من نقد مباشر لافتراض GaitSet القائل بأن ترتيب الإطارات الزمنية غير ضروري وأن المظهر الشكلي للإطار يتضمن ضمنيًا دلالته الزمنية. يجادل مؤلفو GaitPart بأن هذا الافتراض يُخفي حقيقة جوهرية في المشية البشرية، وهي أن أجزاء الجسم المختلفة (الساقان، الجذع، الذراعان) تمتلك أنماط حركة دقيقة (Micro-motions) ومتباينة زمنيًا، لا يمكن تمثيلها بدقة عند التعامل مع الجسم كوحدة واحدة أو عند الاكتفاء بالتجميع الزمني العام. بناءً على ذلك، تقترح الورقة أن كل جزء من الجسم يحتاج إلى نمذجة زمانية مستقلة وقصيرة المدى، وأن الاعتماد على علاقات زمنية طويلة المدى في حركة دورية كالمشية يُعد زائدًا وغير فعّال.



الشكل 17 بنية ومبدأ عمل GAITPART

يوضح الشكل كيف أن التعامل أن النموذج يتعامل مع كل جزء من الجسم كوحدة مستقلة ويراقب تغيراتها ونلاحظ التغيرات ذات التغير الأكبر عند الأطراف منها عند الجذع او الرأس مثلاً.

يتكوّن GaitPart من وحدتين رئيسيتين:

1. Frame-level Part Feature Extractor (FPFE)

شبكة CNN مصممة خصيصًا لاستخراج سمات مكانية على مستوى الأجزاء، تعتمد على طبقة جديدة تسمى Focal Convolution (FConv)، حيث يُقسّم خريطة السمات أفقيًا إلى أجزاء، ويُطبّق الالتفاف بشكل مستقل على كل جزء. تهدف هذه الآلية إلى تقييد مجال الاستقبال (Receptive Field) بحيث يركّز كل جزء على تفاصيله المحلية الدقيقة بدل خلطها مع بقية الجسم.

يوضح الشكل كيف أن النموذج يتعامل مع أجزاء من الإطار الواحد على دفعات.

2. Micro-motion Capture Module (MCM)

لكل جزء من الجسم وحدة MCM مستقلة، تقوم بنمذجة التغيرات الزمنية قصيرة المدى عبر نافذة منزلقة (3 و5 إطارات) باستخدام متوسط وتجميع أعظمي مع آلية انتباه قنوية (Channel-wise Attention). يتم في النهاية تجميع السمات الزمنية لكل جزء باستخدام Temporal Max Pooling، وفق مبدأ أن دورة مشي واحدة كافية لتمثيل النمط الحركي الكامل.

القيمة المضافة مقارنةً بـ GaitSet

يقدّم GaitPart ثلاث إضافات جوهرية للمجال:

1. نمذجة زمنية صريحة ولكن محدودة المدى، تتجنب التعقيد غير الضروري للنماذج العودية أو ثلاثية الأبعاد.
2. تمثيل جزء–مستقل (Part-dependent)، حيث يتعلم كل جزء ديناميكياته الخاصة بدل مشاركة المعاملات عبر الجسم كاملًا.
3. كفاءة أعلى، إذ يبلغ عدد معاملات GaitPart نحو M1.47 فقط، مقارنةً بـ M2.56 في GaitSet، مع أداء أفضل.

مواطن الضعف

رغم تفوقه، لا يخلو GaitPart من قيود بنيوية:

- لا يزال يعتمد كليًا على الصور الظلية الثنائية، مما يجعله حساسًا لأخطاء الاستخلاص (Segmentation Noise).
- التحسين الزمني يظل محليًا، ولا يلتقط العلاقات الطويلة المدى بين الأطوار المختلفة للمشية.
- تقسيم الجسم أفقيًا يفترض محاذاة مثالية للجسم، وهو افتراض قد لا يصمد في البيئات الواقعية غير المضبوطة.
- يعتمد النموذج بدرجة معتبرة على السمات الشكلية الساكنة، وليس الحركة الخالصة فقط.

ان اعتبار الأجزاء المعالجة كل على حدي مستقلة فيما بينها يضعف هذا النموذج وهو ما قد ناقشناه سابقا عند حديثنا عن النماذج المعتمدة على تغير الوضعية أثناء المشي، وبالمقاربة نفسها

نموذج DeepGaitV2: كسر افتراض "النموذج الضحل" في التعرّف على المشية[25]

يمثّل DeepGaitV2 (2023) نقطة تحوّل منهجية في نماذج المشية المعتمدة على الصور الظلية؛ إذ ينطلق من نقد صريح لافتراضٍ ترسّخ بعد GaitSet وGaitPart، مفاده أن المشية مسألة بسيطة نسبيًا تكفيها شبكات CNN ضحلة أو نمذجة زمنية محدودة. تجادل الورقة بأن هذا الافتراض نتج عن الاعتماد المفرط على مجموعات بيانات مخبرية مثل (CASIA-B)، وأن الانتقال إلى بيئات واقعية واسعة النطاق (in-the-wild) يكشف قصور هذه النماذج. بناءً عليه، تقترح DeepGaitV2 أن

تعقيد المشية الواقعية يتطلب تعميق الشبكات ونمذجة زمانية صريحة بدل الاكتفاء بتجميعات زمنية مبسّطة أو حركات دقيقة قصيرة المدى فقط.

البنية المعمارية لنموذج DeepGaitV2

يعتمد نموذج DeepGaitV2 على إطار معماري موحّد للتعرّف على المشية، يتمحور حول تعميق شبكة الاستخلاص وإدخال نمذجة زمنية صريحة، مع الحفاظ على بساطة الإطار العام لتسهيل المقارنة والتعميم. يتكوّن النموذج من ثلاث مراحل رئيسية مترابطة: استخلاص السمات، التجميع المكاني—الزماني، ثم الإسقاط التمييزي.

1. مدخلات النموذج (Input Representation)

يتعامل DeepGaitV2 مع تسلسلات من الصور الظلية الثنائية (Silhouette Sequences) بعد محاذاتها وتوحيد أبعادها (عادةً 64×44). تمثل هذه التسلسلات البعد الزمني للمشية، حيث يُنظر إلى كل إطار على أنه ملاحظة مكانية ضمن حركة دورية.

2. العمود الفقري للاستخلاص (Backbone Network)

يمثّل العمود الفقري قلب النموذج، وقد صُمّم ليكون عميقًا وقابلًا للتدرّج، بخلاف النماذج السابقة التي اكتفت بعدد محدود من الطبقات.

البنية الطبقية

يتكوّن العمود الفقري من:

- طبقة التفاف أولية (Conv Stem).
- أربع مراحل متتالية (Stages).
- كل مرحلة تحتوي على عدد متزايد من الكتل الالتفافية (Residual Blocks)

يتم تعميق الشبكة تدريجيًا (حتى 22 أو 30 طبقة)، مع تقليل الدقة المكانية وزيادة عدد القنوات كلما تعمّقنا، بما يشبه بنى ResNet ولكن مكيّفة خصيصًا لبيانات المشية.

3. نمذجة الزمن: D2 مقابل D3 مقابل P3D

تُعد هذه النقطة الفارقة الأساسية في DeepGaitV2، حيث تُطرح ثلاث صيغ معمارية مختلفة تشترك في الإطار العام وتختلف في كيفية التعامل مع الزمن:

- **DeepGaitV2-2D**
  يستخدم التفافًا ثنائي الأبعاد فقط، ويعامل كل إطار بشكل مستقل مكانيًا، ثم يُدمج البعد الزمني لاحقًا عبر التجميع. يمثّل خط أساس عميق مكانيًا دون نمذجة زمنية صريحة.
- **DeepGaitV2-3D**
  يستخدم التفافًا ثلاثي الأبعاد (D Convolution3) داخل المراحل الوسطى والعميقة، ما يسمح للنموذج بتعلّم التفاعل المكاني—الزماني مباشرة، أي استخراج السمات الحركية من تغير الشكل عبر الزمن.
- **DeepGaitV2-P3D (Pseudo-3D)**
  حلّ وسط معماري يفصل الالتفاف المكاني عن الزمني (D + 1D2)، محققًا توازنًا بين الكلفة الحسابية والقدرة على تمثيل الزمن، مع أداء قريب من النسخة ثلاثية الأبعاد الكاملة.

4. التجميع المكاني—الزماني (Horizontal Pooling & Temporal)

بعد استخراج خرائط السمات، تُطبّق آليتان أساسيتان:

- **Temporal Pooling** لتجميع المعلومات عبر الزمن بعد أن تكون العلاقات الزمنية قد نُمذجت صراحة داخل العمود الفقري.
- **Horizontal Pooling** لتقسيم الجسم إلى شرائط أفقية (على غرار HPM) ، مما يسمح بدمج السمات المحلية والعالمية دون افتراض استقلال صارم للأجزاء كما في GaitPart.

5. رأس الإسقاط التمييزي (Embedding Head)

في المرحلة الأخيرة، تُحوَّل السمات المجمّعة إلى تمثيل تمييزي منخفض الأبعاد باستخدام:

- طبقات Fully Connected.
- BNNeck.
- تدريب قائم على Metric Learning (Triplet Loss).

تهدف هذه المرحلة إلى تعظيم الفصل بين الهويات المختلفة مع الحفاظ على تماسك تمثيلات الشخص نفسه عبر الزوايا والظروف المختلفة.

6. الفلسفة المعمارية العامة

يمكن تلخيص فلسفة DeepGaitV2 المعمارية في ثلاث نقاط:

بعكس الاعتقاد السابق، تُظهر البنية أن نماذج المشية تستفيد بوضوح من الشبكات العميقة وزيادة العمق مفيدة.

الزمن يجب أن يُنمذج صراحة: التجميع وحده غير كافٍ في البيئات الواقعية.

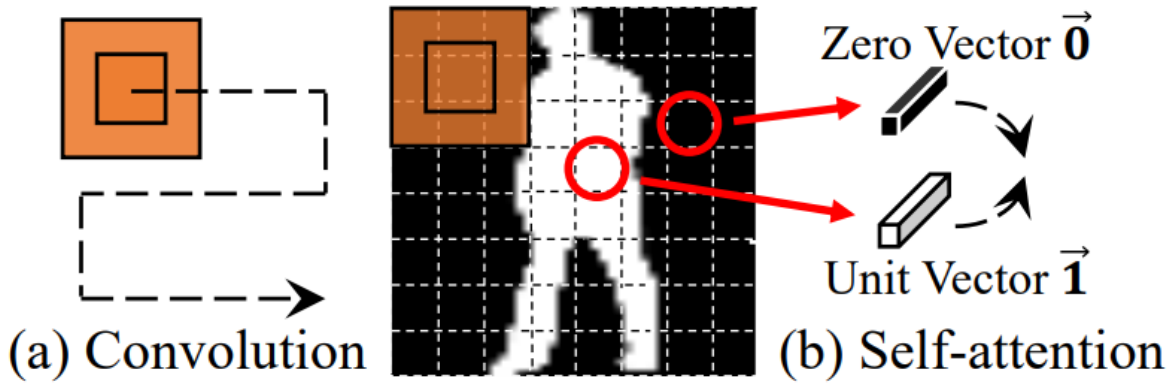
بساطة الإطار، قوة العمق: لم تُستخدم وحدات معقدة أو مخصصة للغاية، بل بُني النموذج على كتل قياسية قابلة للتعميم.

لا يقدّم DeepGaitV2 بنية "ذكية" بالمعنى التقليدي، بل يقدّم بنية صريحة وعميقة تعترف بتعقيد المشية الواقعية. قوته لا تكمن في الحيل المعمارية، بل في كسر افتراضات خاطئة رسّختها البيانات المخبرية، وإعادة تصميم العمود الفقري بما يتناسب مع طبيعة المشكلة.

حدود النماذج الالتفافية وبروز المحوّلات في التعرّف على المشية

تجلى حدود النماذج الالتفافية بصورة أوضح عند تطبيقها في بيئات المراقبة الواقعية، حيث تتسم البيانات بتباين حاد في الزوايا، الإضاءة، جودة الالتقاط، وكثافة الحشود، فضلًا عن عدم انتظام أطوال المقاطع الزمنية وتداخل الحركات بين الأفراد. ففي مثل هذه السيناريوهات، لا تكون السمات المحلية المستخلصة عبر مجالات استقبال محدودة كافية لتمثيل البنية الحركية الكاملة للمشية، حتى عند استخدام شبكات عميقة أو التفافات ثلاثية الأبعاد، إذ يظل تجميع المعلومات العالمية يتم بصورة غير مباشرة ومتأخرة. ويؤدي ذلك في كثير من الأحيان إلى تعلّم أنماط شكلية أو سياقية عارضة بدل التقاط العلاقات الحركية الجوهرية المستقرة عبر الزمن. في المقابل، تتيح نماذج المحوّلات، من خلال آليات الانتباه الذاتي، نمذجة العلاقات طويلة المدى بشكل مباشر، مما يجعلها أكثر قدرة على التعامل مع التشتت الزمني والمكاني الملازم لبيانات المراقبة الواقعية. ومن هذا المنطلق، لا يُعد التحول نحو المحوّلات مجرد تحسين معماري، بل استجابة منهجية لمتطلبات التطبيق العملي في أنظمة المراقبة واسعة النطاق، حيث يصبح التعميم والاستقرار عبر الظروف المتغيرة عاملين حاسمين في تقييم فعالية نماذج التعرّف على المشية.

تقوم بنى المحوّلات—وعلى وجه الخصوص SwinGait [25] يربط الكتل (patches) ضمن سياقات أكثر متانة وشمولًا مقارنةً بالشبكات الالتفافية، إذ لا تقتصر عملية التعلّم على مجالات استقبال محلية ثابتة، بل تعتمد على نمذجة علائقية صريحة بين الرقع عبر آليات الانتباه الذاتي. غير أن تطبيق المحوّلات على الصور الظلية يواجه تحديًا خاصًا يُعرف بمشكلة dumb patches، الناتجة عن الطبيعة الثنائية لهذا التمثيل؛



الشكل 18 الرقع عديمة القيمة في الصور الظلية

فمعظم الرقع المستخرجة من الصورة تكون إما سوداء بالكامل أو بيضاء بالكامل، ولا تحمل في حد ذاتها أي معلومات تمييزية، بينما تتركز المعلومات الحركية والشكلية ذات القيمة فقط في الرقع الواقعة على حدود الجسم (Body Contours). في المحوّلات التقليدية ذات الرقع الثابتة، يؤدي هذا الأمر إلى هيمنة رقع عديمة الدلالة على عملية الانتباه، مما يضعف قدرة النموذج على تعلّم علاقات ذات معنى. هنا تبرز أهمية التصميم الهرمي في SwinGait، حيث يُطبّق الانتباه داخل نوافذ محلية صغيرة تُزاح (Shifted Windows) بين الطبقات المتعاقبة، ما يسمح بإعادة توزيع الرقع الحدّية عبر سياقات مختلفة. ونتيجة لعملية الإزاحة، قد تنتقل رقعة تقع على هامش نافذة في طبقة ما لتصبح في مركز نافذة أخرى في طبقة لاحقة، مما يمنحها وزنًا سياقيًا أعلى ويُبرز أهميتها العلائقية. ويتعزز هذا الأثر أكثر من خلال التمثيل الهرمي، حيث تُدمج الرقع تدريجيًا في مستويات أعلى، فتتحول المعلومات الحدّية المحلية إلى سمات عالمية أكثر استقرارًا ودلالة. وبذلك، لا يكتفي SwinGait بتخفيف أثر الرقع غير المهمة، بل يعيد تنظيم الرقع المفيدة ضمن سياقات متعددة المستويات، ما يجعله أكثر قدرة على استخراج تمثيلات حركية متماسكة من بيانات ظلية فقيرة دلاليًا. وتكمن أهمية SwinGait تحديدًا في كونه يقدّم حلًا معماريًا يتعامل بوعي مع قيود التمثيل الثنائي، ويحوّل ضعف الصور الظلية—أي فقرها المعلوماتي—إلى دافع لاعتماد نمذجة علائقية هرمية قادرة على دعم التعرّف على المشية في البيئات الواقعية المعقّدة.

مقارنة بين النماذج المذكورة أعلاه من ناحية Rank-1:

النموذج	سنة النشر	CASIA-B (NM)	CASIA-B (BG)	CASIA-B (BG)
GaitSet	2019	95	87.2	70
GaitPart	2020	96	91.5	79
DeepGaitV2-2D	2023	91	88	75
DeepGaitV2-P3D	2023	92	90	78
DeepGaitV2-3D	2023	92	91	79
SwinGait	2023	96	93	84

الجدول 9 مقارنة النماذج المعتمدة على الصور الظلية على مجموعة بيانات CASIA-B

النموذج	سنة النشر	OU-MVLP	Gait3D	GREW
GaitSet	2019	87	36	46
GaitPart	2020	89	45	60
DeepGaitV2-2D	2023	86	68	72
DeepGaitV2-P3D	2023	88	74	77

79	72	89	2023	DeepGaitV2-3D
82	76	90	2023	SwinGait

الجدول 10 الجدول 9 مقارنة النماذج المعتمدة على الصور الظلية على مجموعات بيانات المشية

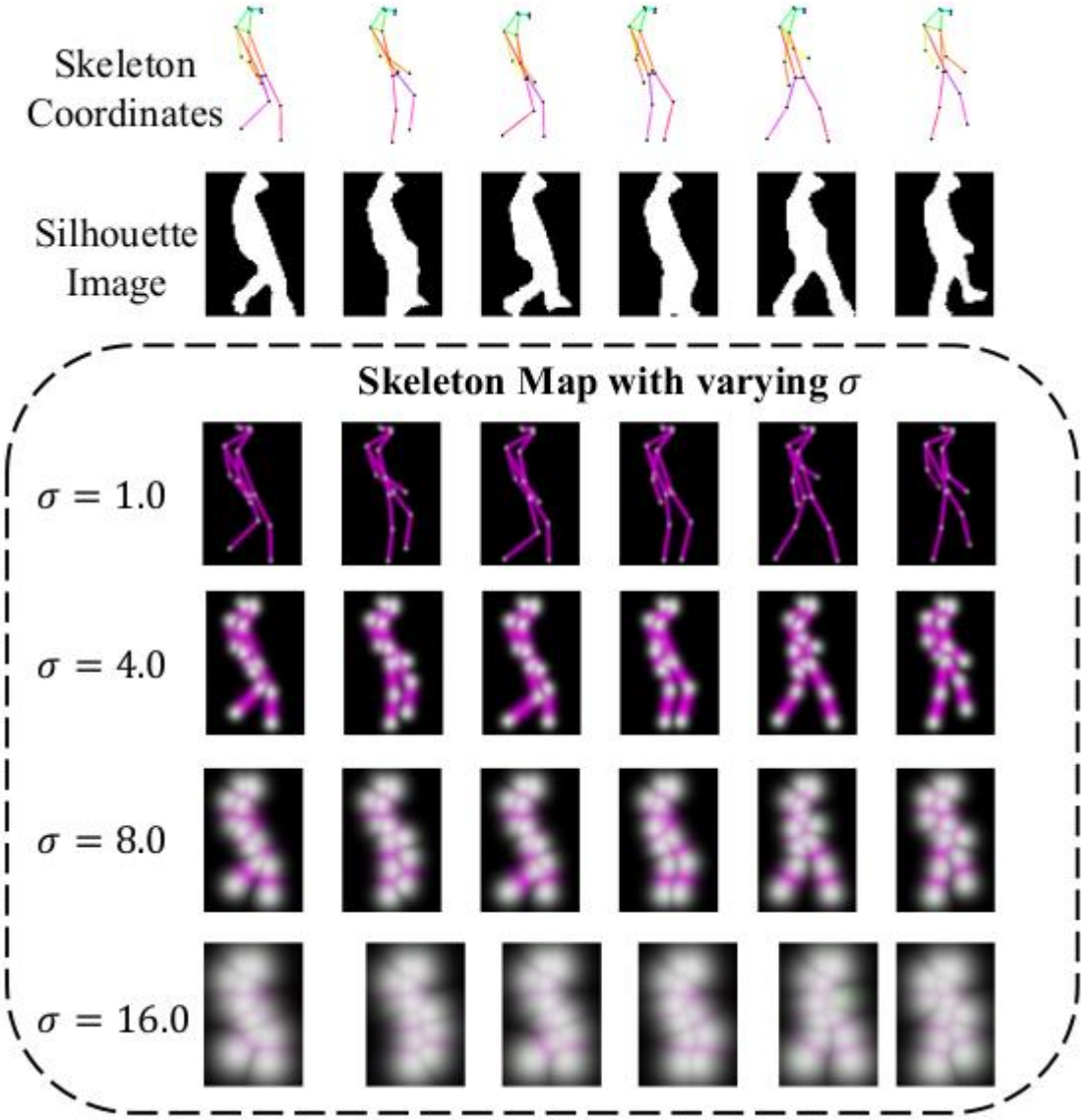
تُظهر نتائج الجدولين أن الفروق بين النماذج تكون محدودة نسبيًا على مجموعات البيانات المخبرية مثل CASIA-B، حيث تحقق معظم النماذج دقة مرتفعة في حالة المشي الطبيعي، ما يجعل التفوق العددي أقل دلالة على القوة التعميمية الفعلية. في المقابل، تتجلى الفروق الحقيقية عند الانتقال إلى مجموعات بيانات واقعية مثل Gait3D وGREW، إذ ينخفض أداء النماذج المعتمدة على التجميع الزمني أو النمذجة الجزئية البسيطة بشكل حاد، كما في GaitSet وGaitPart، بينما تحقق نماذج DeepGaitV2 تحسنًا واضحًا بفضل تعميق الشبكات ونمذجة الزمن بشكل صريح. ويبرز SwinGait بوصفه النموذج الأكثر قدرة على التعميم، إذ يتفوّق على جميع النماذج السابقة في البيئات الواقعية، مما يؤكد أن الانتقال من المعالجة المحلية للشبكات الالتفافية إلى النمذجة العلائقية الهرمية القائمة على الانتباه الذاتي يمثل تحولًا منهجيًا ضروريًا للتعرّف العملي على المشية في أنظمة المراقبة الواقعية واسعة النطاق.

3.4.4. النماذج الهجينة في التعرّف على المشية

إلى جانب النماذج المعتمدة حصريًا على الصور الظلية أو الهياكل والوضعيات، برزت في السنوات الأخيرة فئة من النماذج الهجينة التي تسعى إلى الجمع بين مزايا تمثيلات مختلفة للتغلب على القيود البنيوية لكل تمثيل على حدة. تقوم هذه النماذج إما بدمج مصادر متعددة للمعلومة (مثل الشكل الظلي، الهيكل العظمي، أو السمات الحركية)، أو بإعادة صياغة تمثيل بنيوي صرف ضمن صيغة صورية قابلة للمعالجة بواسطة الشبكات العميقة المصممة للصور. ويهدف هذا الاتجاه إلى تحقيق توازن بين نقاء التمثيل الحركي الذي توفره الهياكل والوضعيات، والقوة التمييزية والاستقرار المعماري الذي توفره النماذج الصورية العميقة، خاصة في البيئات الواقعية حيث تتأثر الصور الظلية بالضجيج، بينما تعاني الهياكل من عدم الاكتمال أو أخطاء الاستخلاص. وبهذا المعنى، لا تُعد النماذج الهجينة مجرد دمج تقني، بل تمثل توجهًا منهجيًا لإعادة تعريف تمثيل المشية بما يخدم متطلبات التعميم العملي.

مثالاً عنها نذكر النموذج المنشور عام 2024 SkeletonGait [26] وامتداده SkeletonGait++ بوصفهما مقاربة جديدة للتعرّف على المشية تعتمد على **التمثيل الهيكلي الخالص** مع إعادة صياغته ضمن شكل **تمثيل صوري شبكي** (Skeleton Map). ينطلق العمل من ملاحظة أن نماذج الصور الظلية، رغم فعاليتها، تعاني من حساسية عالية لتغيرات المظهر

والملابس، بينما تواجه نماذج الهياكل التقليدية القائمة على الرسوم البيانية تحديات تتعلق بتعقيد البنية وصعوبة الاستفادة من الشبكات العميقة المصممة للصور. لمعالجة ذلك، تقترح الورقة إسقاط إحداثيات المفاصل ثنائية الأبعاد إلى خرائط حرارية متعددة القنوات تحافظ على العلاقات البنيوية بين المفاصل دون تضمين أي معلومات شكلية أو مظهرية، ثم معالجتها باستخدام عمود فقري عميق مشتق من DeepGaitV2. كما هي موضحة في الشكل التالي ويتم إنشاؤها بقياس الفضاء الاحتمالي لوقوع كل من العقد والوصلات في البيان الممثل للوضعية



الشكل 19 الهيكل والوضعية مبنية باستخدام الخرائط الحرارية وفقاً لمقاربة SKELETONGAIT

أظهرت النتائج التجريبية، على مجموعات بيانات مثل CASIA-B وGREW، أن النموذج المقترح يحقق أداءً تنافسيًا بل ومتفوّقًا في بعض السيناريوهات مقارنةً بنماذج الصور الظلية، مع استقرار أعلى تجاه تغيرات المظهر. وتخلص الورقة إلى أن هذا النوع من التمثيلات الهجينة—الهيكلية من حيث المحتوى والصورية من حيث المعالجة—يمثل اتجاهًا واعدًا لسد الفجوة بين نقاء التمثيل الحركي وقابلية الاستفادة من البنى العميقة الحديثة، خصوصًا في سياق أنظمة المراقبة الواقعية.

تُبرز المقارنة بين SkeletonGait وكلٍ من DeepGaitV2 وSwinGait ثلاثة مسارات منهجية مختلفة للتعرّف على المشية. يعتمد DeepGaitV2 على تعميق الشبكات الالتفافية ونمذجة الزمن بشكل صريح لمعالجة الصور الظلية، وقد حسّن الأداء في البيئات الواقعية مقارنةً بالنماذج الضحلة، إلا أنه يبقى حساسًا لتغيرات المظهر وجودة استخراج الصور الظلية. من جانبه، يتجاوز SwinGait حدود الالتفاف عبر الانتباه الذاتي الهرمي، ما يمنحه قدرة أعلى على نمذجة العلاقات طويلة المدى وتحقيق تفوق واضح على بيانات المراقبة الواقعية، وإن كان ذلك على حساب تعقيد معماري وكلفة حسابية أعلى، مع استمرار الاعتماد على تمثيل ظلي فقير دلاليًا. في المقابل، يعالج SkeletonGait المشكلة على مستوى التمثيل، إذ يعتمد تمثيلًا هيكليًا خالصًا يقلل الحساسية للمظهر، لكنه يصبح مرتبطًا مباشرة بجودة تقدير الوضعيات. وتشير هذه المقارنة إلى أن التقدم الفعلي في التعرّف على المشية ينبع من تفاعلٍ متوازن بين اختيار التمثيل، ونمذجة الزمن، والبنى العلائقية الداعمة للتعميم الواقعي

الفرق بين SkeletonGait و SkeletonGait++

يمثّل كل من SkeletonGait و SkeletonGait++ إطارين متتابعين ضمن المقاربة نفسها، ويشتركان في الفلسفة العامة القائمة على **التمثيل الهيكلي الخالص** للمشية مع إعادة صياغته ضمن هيئة **خرائط صورية** (Skeleton Maps)، إلا أن النسخة المحسّنة SkeletonGait++ تقدّم تحسينات منهجية على مستوى التمثيل، والدمج الزمني، واستراتيجية التعلّم.

مقارنة النتائج Rank-1%:

النموذج	CASIA-B (NM)	CASIA-B (BG)	CASIA-B (CL)
DeepGaitV2-3D	92	91	79
SwinGait	96	93	84

84	93	96	SkeletonGait
86	94	97	SkeletonGait++

الجدول 11 مقارنة SKELETONGAIT مع نماذج الصور الظلية على محموعة البيانات CASIA-B

GREW	النموذج
79	DeepGaitV2-3D
83	SwinGait
81	SkeletonGait
84	SkeletonGait++

الجدول 12 مقارنة SKELETONGAIT مع نماذج الصور الظلية على محموعة البياناتGREW

يحقق SkeletonGait أداءً تنافسيًا جدًا مع SwinGait رغم اعتماده على تمثيل هيكلي خالص دون أي معلومات شكلية.

يتفوّق SkeletonGait++ على SwinGait على GREW ويضاهيه على CASIA-B، ما يؤكد قوة التمثيل الهيكلي في البيئات الواقعية.

تثبت النتائج أن **تحسين التمثيل** قد يكون بنفس أهمية (وأحيانًا أهم من) تعقيد البنية المعمارية.

تفوق SkeletonGait++ على DeepGaitV2 يبرز حدود الصور الظلية حتى مع النمذجة الزمنية العميقة.

3.5. خصائص الهيئة والمظهر (Appearance) في مسألة إعادة التعرف

تُعدّ النماذج المعتمدة على المظهر (Appearance-Based Models) النهج الكلاسيكي والأكثر شيوعًا في أنظمة إعادة التعرّف على الأشخاص. وتعتمد هذه النماذج على استخلاص السمات البصرية مباشرةً من صور RGB، بهدف مطابقة الشخص نفسه عبر كاميرات غير متداخلة، رغم اختلاف زاوية التصوير، والإضاءة، والخلفية، ووجود الحجب الجزئي.

1. المبدأ الأساسي

تعتمد النماذج المعتمدة على المظهر على تعلّم فضاء تمثيلي تمييزي تكون فيه صور الهوية نفسها متقاربة، بينما تكون صور الهويات المختلفة متباعدة. ويُستخرج هذا التمثيل اعتمادًا على المظهر البصري العام والمحلي، بما في ذلك:

- لون الملابس وملمسها.
- شكل الجسم وحدود الظل الخارجي (بصورة ضمنية وغير صريحة).
- الإكسسوارات مثل الحقائب وحقائب الظهر والقبعات.
- التفاصيل الدقيقة مثل الشعارات، والخطوط، والزخارف.

ويُنظر إلى كل صورة (أو مقطع قصير من الإطارات) باعتبارها ملاحظة بصرية ساكنة، دون نمذجة صريحة للحركة البشرية أو الخصائص الحيوية للمشي.

2. البنية المعمارية النموذجية

تتبع معظم أنظمة ReID المعتمدة على المظهر بنية معمارية متقاربة، تتضمن المراحل الآتية:

1. **العمود الفقري (Backbone) – شبكة التفاف عميقة أو محوّل بصري**
   - لاستخلاص السمات البصرية عالية المستوى.
   - يشمل ذلك الشبكات الالتفافية العميقة أو المحوّلات البصرية.
2. **تجميع السمات (Feature Aggregation)**
   - تجميع عالمي (متوسط أو أقصى تجميع).
   - أو تجميع قائم على الأجزاء (تقسيم الجسم إلى شرائط أفقية أو مناطق اهتمام).
3. **رأس التمثيل (Embedding Head)**
   - لإنتاج متجه تمثيل مضغوط للهوية.
   - غالبًا مع طبقات تطبيع لتحسين الاستقرار التمييزي.
4. **هدف التعلّم القياسي (Metric Learning Objective)**
   - لضمان فصل الهويات في فضاء السمات.

وتؤدي هذه البنية دور بصمة بصرية تمثل هوية الشخص.

وتأكيداً على شيوع هذه المقاربة سواء على صعيد اعتماد المظهر كخاصية مميزة أو من ناحية اعتماد المعمارية السابقة، نجد فريقاً من الباحثين قام بإنشاء إطار عملي باسم torechreid وهو معتمد على مكتبة PyTorch الشهيرة.

3. نقاط القوة في النماذج المعتمدة على المظهر

لا تزال هذه النماذج مهيمنة في التطبيقات الواقعية لما تتمتع به من مزايا، منها:

- **قدرة تمييزية عالية** في السيناريوهات قصيرة المدى.
- **عدم الحاجة إلى تسلسل زمني**؛ تعمل على صورة واحدة.
- **توفر مجموعات بيانات واسعة** مبنية على صور RGB.
- **سهولة الدمج والتطبيق** ضمن أنظمة المراقبة القائمة.

وفي البيئات المضبوطة أو شبه المضبوطة، تحقق هذه النماذج عادةً أداءً متقدمًا في مهام الاسترجاع.

4. القيود البنيوية الأساسية

على الرغم من نجاحها، تعاني النماذج المعتمدة على المظهر من قيود جوهرية:

أ) الحساسية لتغيّر الملابس

نظرًا لاعتمادها الكبير على الملابس، يتدهور الأداء بشكل ملحوظ عند:

- تغيير الشخص لملابسه.
- ارتداء زي موحّد من قب عدة أشخاص.
- حدوث تغيّرات موسمية.

وهنا تتميز النماذج المعتمدة على المشية إذ أنها تتعامل مع معطيات غير مرتبطة باللون من الأساس.

ب) الانحياز لزاوية التصوير والحجب

قد تهيمن حقائب الظهر أو الزوايا الجانبية أو الحجب الجزئي على التمثيل، ما يؤدي إلى مطابقات خاطئة.

ج) غياب الدلالة الزمنية

هذه النماذج:

- لا ترمز أسلوب المشي أو أنماط الحركة.
- لا تستفيد من انتظام الحركة عبر الزمن.
- تتعامل مع الإطارات المتتابعة كعينات مستقلة.

وهو ما يميزها جذريًا عن النماذج المعتمدة على المشية أو الوضعية.

إطارTorchReID وفلسفته المنهجية في أنظمة إعادة التعرّف على الأشخاص

يُعدّ **TorchReID** أحد الأطر البرمجية البحثية المرجعية في مجال إعادة التعرّف على الأشخاص (Person Re-Identification)، وقد صُمِّم انطلاقًا من رؤية منهجية تهدف إلى توحيد البنية التجريبية لنماذج إعادة التعرّف المعتمدة على المظهر البصري، مع ضمان القابلية لإعادة الإنتاج، وسهولة المقارنة العلمية، والفصل الصارم بين مكوّنات النظام المختلفة.

الفلسفة العامة للإطار

تنطلق فلسفة TorchReID من افتراض علمي مفاده أن مشكلة إعادة التعرّف على الأشخاص لا ينبغي التعامل معها كنظام مغلق أو نموذج أحادي، بل كـ منظومة معيارية (Modular System) يمكن تفكيكها إلى وحدات مستقلة وظيفيًا، تشمل البيانات، والنماذج، ودوال الخسارة، وآليات التدريب، وبروتوكولات التقييم. ويتيح هذا التفكيك دراسة أثر كل مكوّن على حدة، بما ينسجم مع متطلبات البحث العلمي المنضبط.

وبناءً على ذلك، لا يهدف TorchReID إلى تقديم نموذج أمثل بعينه، وإنما إلى توفير بيئة تجريبية موحّدة تُستخدم لتقييم ومقارنة مختلف النماذج والخوارزميات ضمن شروط متكافئة.

البنية المفاهيمية لإطار TorchReID

يمكن توصيف بنية TorchReID على أنها تتألف من خمس طبقات مفاهيمية مترابطة، تمثل معًا الهيكل المنهجي للنظام.

أولًا: طبقة إدارة البيانات

تُعنى هذه الطبقة بتنظيم وتحميل مجموعات بيانات إعادة التعرّف، مع توحيد آلية التعامل مع اختلاف بنيتها الداخلية. وتشمل وظائفها فصل البيانات إلى مجموعات التدريب، والاستعلام، والمعرض، إضافةً إلى إدارة معرفات الأشخاص ومعرفات الكاميرات.

وتكمن أهمية هذه الطبقة في كونها تفصل النموذج عن مصدر البيانات، مما يسمح بتطبيق النموذج نفسه على مجموعات بيانات متعددة، أو إجراء تقييمات عابرة للمجموعات (Cross-Dataset Evaluation)، دون الحاجة إلى تعديل البنية البرمجية.

ثانيًا: طبقة النماذج التمثيلية

يعامل TorchReID نموذج إعادة التعرّف بوصفه مستخرجًا للسمات التمييزية، لا مجرد مصنّف تقليدي. ويتكوّن النموذج عادةً من عمود فقري بصري مسؤول عن استخلاص السمات عالية المستوى، يتبعه رأس تمثيلي يُنتج شعاع مضغوطًا يمثّل هوية الشخص.

ويعكس هذا التصميم فهمًا مفاده أن جوهر المشكلة لا يكمن في إسناد تسمية فئوية، بل في بناء فضاء تمثيلي تكون فيه المسافات الهندسية معبّرة عن درجة التشابه الهويّاتي بين العينات.

ثالثًا: طبقة دوال الخسارة

يعتمد TorchReID على مزيج من دوال الخسارة التصنيفية والقياسية، بما ينسجم مع طبيعة إعادة التعرّف بوصفها مسألة تعلّم قياسي (Metric Learning). وتهدف هذه الدوال إلى تقليص التباعد بين تمثيلات العينات العائدة للشخص نفسه، وزيادة التباعد بين تمثيلات الأشخاص المختلفين.

وتُجسّد هذه المقاربة تصورًا هندسيًا للهوية، حيث تُتعلّم الهوية بوصفها موضعًا مستقرًا ضمن فضاء السمات، لا مجرد وسم رقمي.

رابعًا: طبقة محرّك التدريب والتقييم

تفصل هذه الطبقة منطق التدريب والتنفيذ عن تفاصيل النموذج ذاته. فهي تتولى إدارة الحلقات التدريبية، وتحديث الأوزان، وتنفيذ التقييم الدوري، مع دعم التدريب المتوازي على وحدات معالجة متعددة.

ويُسهم هذا الفصل في تعزيز قابلية التوسّع، ويُيسّر إجراء التجارب المقارنة دون إعادة كتابة الشيفرة الخاصة بالتدريب.

خامسًا: طبقة التقييم والمقاييس

يعتمد TorchReID مقاييس تقييم تتوافق مع طبيعة المهمة بوصفها نظام استرجاع، لا نظام تصنيف مغلق. وتشمل هذه المقاييس ترتيب الاسترجاع (Rank-k) ومتوسط الدقة (mAP)، والتي تعكس قدرة النموذج على استرجاع الهوية الصحيحة ضمن سياق متعدد الكاميرات.

وتُعد هذه المقاييس أكثر تعبيرًا عن الأداء الواقعي لأنظمة إعادة التعرّف مقارنةً بمقاييس الدقة التقليدية.

الأهمية البحثية لإطار TorchReID

تنبع أهمية TorchReID من كونه:

- يوفّر خط أساس معياري معتمد على نطاق واسع،
- يقلّل من التحيّزات البرمجية بين الدراسات المختلفة،
- يسهّل إعادة إنتاج النتائج والتحقق منها،
- ويدعم التحليل المقارن المنهجي بين النماذج والخوارزميات.

ولهذا، تُستخدم نتائجه غالبًا كمرجع ضمني عند تقييم النماذج المعتمدة على المظهر في الأدبيات الحديثة.

النماذج المعتمدة على المظهر

إن التعامل مع مسألة إعادة المطابقة مع الصور بشكلها المباشر يقترح بداهة استخدام الشبكات العصبونية والنماذج التي يشيع استخدامها مع الصور بشكل عام ولعل من أشهرها النماذج المبنية على معمارية resnet وهو ما جعل العديد من الأبحاث تطرح ما يسمى ب resnet basline models تختلف في عواملها الفائقة وأبعادها وما شابه ذلك.

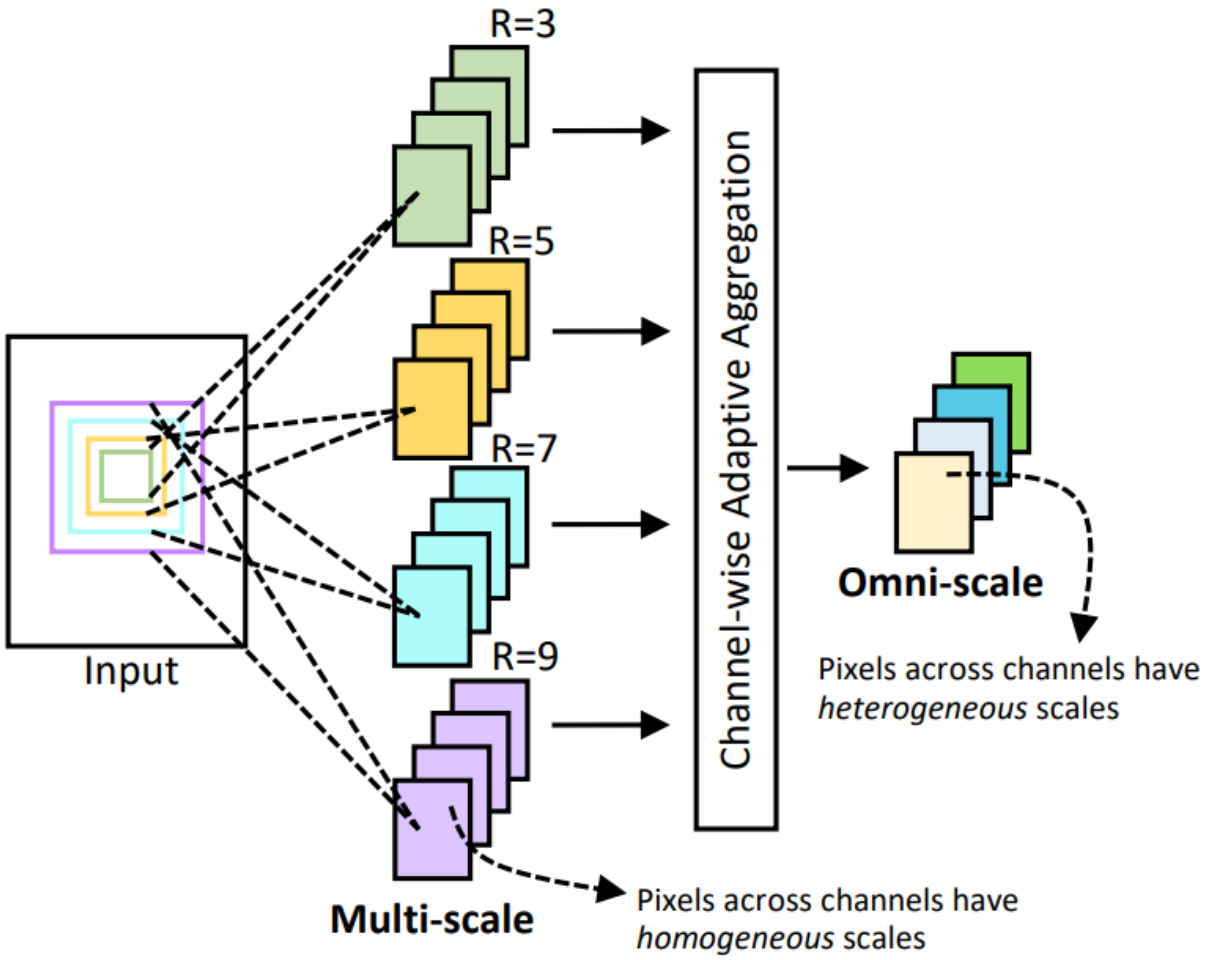
لذلك سنركز على مقاربة مختلفة وهي مقاربة نموذج Omni-Scale Network او اختصاراً OSNet.

قدّم OSNet مقاربة معمارية مختلفة تقوم على مفهوم التعلّم الشامل للمقاييس (Omni-Scale Feature Learning) ، حيث تنطلق من فرضية أن السمات التمييزية في إعادة التعرّف لا تكون محصورة في تفاصيل محلية دقيقة أو في بنية جسدية عامة فحسب، بل في تركيبات دلالية تجمع بين مقاييس متجانسة وغير متجانسة. ويُعد هذا الطرح تحوّلًا مفاهيميًا مقارنةً بالأعمال السابقة التي اكتفت إمّا بتجميع سمات من طبقات مختلفة، أو باستخدام مسارات متعددة ذات مقاييس ثابتة.

تعتمد معمارية OSNet على كتل متبقية متعددة المسارات، بحيث يختص كل مسار باستخلاص سمات عند مجال استقبالي مختلف داخل الطبقة نفسها. وتمتاز هذه البنية بوجود بوابة دمج موحّدة وديناميكية تقوم بتجميع السمات متعددة المقاييس على مستوى القنوات، وبطريقة معتمدة على دخل الصورة، بدلًا من الدمج الثابت أو الخشن المستخدم في نماذج أقدم. ويتيح

هذا التصميم للنموذج أن ينتج، عند الحاجة، تمثيلات أحادية المقياس، أو تمثيلات هجينة تجمع تفاصيل محلية مع سياق مكاني أوسع، وهو ما يتلاءم مع الطبيعة الدقيقة لمهمة إعادة التعرّف.

إلى جانب الإسهام المفاهيمي، يتميّز OSNet بخفة معمارية ملحوظة، إذ يعتمد على التلافيف القابلة للفصل (Depthwise Separable Convolutions)، ما يقلّل عدد المعاملات بصورة كبيرة مقارنةً بالنماذج المعتمدة على ResNet-50، مع الحفاظ على قدرة تمثيلية عالية. وقد أظهرت النتائج التجريبية تفوّق OSNet على عدد كبير من النماذج الأثقل وزنًا عبر عدة مجموعات بيانات معيارية، محققًا أداءً متقدمًا من حيث Rank-1 وmAP، حتى عند تدريبه من الصفر على مجموعات بيانات ذات حجم متوسط



الشكل 20 مبدأ عمل OMNIS-SCALE NETWORK

وعلى الرغم من هذه المزايا، يظل OSNet، شأنه شأن بقية النماذج المعتمدة على المظهر، محدودًا بكونه لا يستفيد من المعلومات الزمنية أو الحركية، ولا يُنمذج البنية الهيكلية للجسم أو أسلوب المشي. ويجعل ذلك أداءه عرضة للتراجع في سيناريوهات تغيّر

الملابس أو إعادة التعرّف طويلة الأمد. ولهذا، تشير الاتجاهات البحثية الحديثة إلى استخدام OSNet كنواة مظهرية قوية ضمن أنظمة هجينة، تُدمج فيها سمات الوضعية أو المشية أو التمثيلات الدلالية لتعويض هذه الحدود البنيوية.

بناءً عليه، يمكن اعتبار OSNet أحد أكثر النماذج نضجًا ضمن فئة النماذج المعتمدة على المظهر، لما يقدّمه من توازن بين العمق المفاهيمي والكفاءة العملية، فضلًا عن كونه مرجعًا معياريًا تُقاس عليه المقاربات اللاحقة في أدبيات إعادة التعرّف على الأشخاص.

دور المحوّلات والمحوّلات البصرية كنقطة انتقال في تطوّر نماذج ReID

مثّل OSNet نقطة اكتمال منهجية لمسار النماذج المعتمدة على المظهر داخل الأطر الالتفافية، حيث بلغ تحسين التمثيل البصري متعدد المقاييس داخل CNN حدًّا ناضجًا كشف بوضوح أن التقدّم اللاحق لم يعد ممكنًا عبر تعديلات معمارية طفيفة. ونتيجةً لذلك، تحوّل سؤال البحث من *كيفية تحسين المظهر* إلى *كيفية تجاوز حدوده البنيوية*. في هذا السياق، برزت المحوّلات البصرية (Vision Transformers) بوصفها مرحلة انتقالية مفصلية، لا بوصفها بديلًا نهائيًا لـCNN ، بل كآلية كشفت في آنٍ واحد حدود الشبكات الالتفافية وحدود المظهر ذاته.

أظهرت ورقة Vision Transformer (ViT) أن آلية الانتباه الذاتي قادرة على نمذجة العلاقات بعيدة المدى بين أجزاء الصورة بصورة تتجاوز القيود المحلية لـCNN ، وهو ما مثّل قفزة مفاهيمية مقارنةً بنماذج مثل OSNet. غير أن هذا التفوق كان مشروطًا بكلفة حسابية مرتفعة وحاجة إلى بيانات ضخمة، إضافةً إلى غياب التحيّزات الاستقرائية والبنية الهرمية التي أثبتت أهميتها في المهام البصرية الواقعية. وقد أكّد ذلك أن المحوّلات الخالصة، رغم قوتها التمثيلية، ليست حلًا عمليًا قائمًا بذاته في أنظمة ReID.

استجابةً لهذه القيود، جاءت Swin Transformer لتعيد إدخال مفاهيم المحلية، والتدرّج المكاني، وتعدّد المقاييس ضمن إطار تحويلي، عبر تقييد الانتباه الذاتي داخل نوافذ محلية متحركة وبناء تمثيل هرمي قريب من فلسفة CNN. ويعكس هذا التطوّر إدراكًا بحثيًا بأن المحوّلات لا تُقصي الشبكات الالتفافية، بل تتقاطع معها، وتؤدي دورًا وسيطًا في إعادة توجيه مسار البحث من تحسين المظهر إلى إعادة التفكير في طبيعة التمثيل الهويّاتي ذاته.

ضمن مجال إعادة التعرّف على الأشخاص، تؤكّد هذه المرحلة الانتقالية أن ما بعد OSNet لم يكن انتقالًا خطيًا نحو معماريات أدقّ، بل تحوّلًا مفاهيميًا كشف محدودية الاعتماد على الصورة الثابتة وحدها، سواء أكانت ممثلة عبر CNN أو عبر محوّلات بصرية. ونتيجةً لذلك، اتجهت الأبحاث اللاحقة إلى نماذج هجينة وخفيفة تدمج CNN والمحوّلات انتقائيًا، ثم إلى مسارات تتجاوز المظهر نحو دمج البنية الجسدية والديناميكيات الحركية والتمثيلات الدلالية. وعليه، يمكن النظر إلى المحوّلات البصرية

بوصفها حلقة وصل حاسمة بين اكتمال نماذج المظهر (كما في OSNet) وبين الأنظمة المركّبة ومتعددة الإشارات التي تميّز المشهد البحثي المعاصر في ReID.

النماذج الهجينة الكفوءة:(2025) TE-TransReID [27] مثالاً

بعد أن بلغ مسار النماذج المعتمدة على المظهر داخل الأطر الالتفافية ذروته مع OSNet، حيث تم تحسين التمثيل متعدد المقاييس بكفاءة حسابية عالية، اتجهت الأبحاث اللاحقة إلى المحوّلات البصرية بوصفها وسيلة لتجاوز القيود المحلية لـCNN ونمذجة العلاقات المكانية بعيدة المدى. وقد أثبتت نماذج مثل TransReID جدوى هذا التوجّه من حيث الدقة، إلا أنها كشفت في الوقت ذاته عن عبء حسابي مرتفع وحساسية عالية لحجم البيانات، مما حدّ من قابليتها للتطبيق العملي في أنظمة المراقبة الواقعية. في هذا السياق، يُقترح نموذج TE-TransReID بوصفه حلًا وسيطًا ناضجًا يوازن بين القدرة التمثيلية للمحوّلات وكفاءة الشبكات الالتفافية، دون الانخراط في التضخيم المعماري الذي ميّز الجيل الأول من نماذج Transformer-based ReID.

تعتمد فلسفة TE-TransReID على التكامل الانتقائي بين تمثيلات محلية وعالمية، حيث تُستخرج السمات المحلية باستخدام شبكة CNN خفيفة (MobileNetV2) متخصصة في التقاط التفاصيل الدقيقة، في حين تُستخلص السمات العالمية عبر جزء مقتطع من Vision Transformer، يقتصر على عدد محدود من الطبقات الأولى لتقليل الكلفة الحسابية. ويتميّز النموذج بجعل عملية الدمج عنصرًا بنيويًا أساسيًا، عبر وحدات دمج كفوءة تعمل على مستويي المتجهات والرقع، ما يسمح بتعلّم مساهمة كل من السمات المحلية والعالمية بصورة ديناميكية. وقد أظهرت النتائج التجريبية أن هذا التصميم يحقق أداءً منافسًا للنماذج التحويلية الثقيلة، مع تقليل ملحوظ في عدد المعاملات والتعقيد الحسابي.

من منظور تطوّر المجال، يمثّل TE-TransReID خلاصة مرحلة ما بعد OSNet: فهو يؤكّد أن المحوّلات البصرية ليست بديلًا مباشرًا للشبكات الالتفافية، بل مكوّنًا مكمّلًا لها ضمن بنى هجينة مدروسة. كما يرسّخ فكرة أن التقدّم في ReID لم يعد مرهونًا بزيادة عمق النماذج أو تعقيدها، بل بإعادة توزيع الأدوار بين المكوّنات المعمارية لتحقيق توازن عملي بين الدقة والكفاءة. ومع ذلك، يظل النموذج محصورًا ضمن إطار المظهر البصري، مما يدعم الاتجاهات البحثية اللاحقة التي تسعى إلى تجاوز الصورة الثابتة نحو دمج البنية الجسدية والديناميكيات الحركية في أنظمة ReID متعددة الإشارات

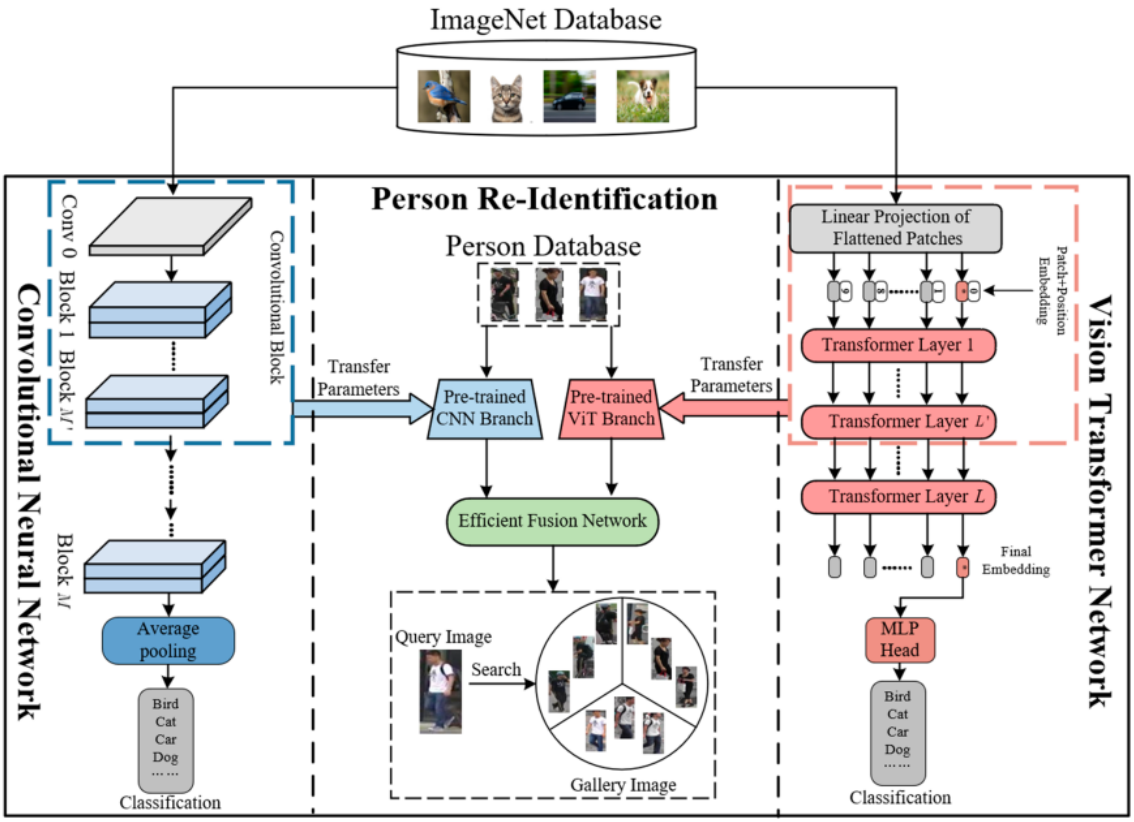
المعمارية الهجينة وشرح مفهوم الاقتطاع في TE-TransReID

تعتمد معمارية نموذج TE-TransReID على تصميم هجيني ثنائي الفروع يهدف إلى تحقيق توازن عملي بين القدرة التمثيلية والكلفة الحسابية، وذلك عبر الجمع الانتقائي بين الشبكات الالتفافية والمحوّلات البصرية. يتكوّن النموذج من فرع محلي قائم على شبكة التفاف خفيفة، مسؤول عن استخلاص السمات الدقيقة المرتبطة بالمظهر الخارجي، مثل تفاصيل الملابس وبنية الأطراف، ومن فرع عالمي قائم على محوّل بصري، يُستخدم بصورة محدودة لتوفير وعي سياقي مكاني واسع. ويُدمج ناتج الفرعين عبر وحدات دمج ديناميكية تعمل على مستويات مختلفة، بما يسمح ببناء تمثيل موحّد يجمع بين الدقة المحلية والسياق العالمي دون تضخيم معماري.

في هذا الإطار، يُقصد بمفهوم Vision Transformer المقتطع استخدام عدد محدود من الطبقات التحويلية الأولى فقط من نموذج Vision Transformer القياسي، مع استبعاد الطبقات العميقة اللاحقة. ويهدف هذا الاقتطاع إلى الاستفادة من قدرة الانتباه الذاتي على نمذجة العلاقات المكانية بعيدة المدى، مع تجنّب الكلفة الحسابية المرتفعة وخطر فرط التكيّف المرتبطين باستخدام المحوّل كامل العمق. وتُوظَّف الطبقات الأولى للمحوّل بوصفها وحدة نمذجة سياقية عامة، في حين تُسند مهمة التمييز الدقيق واستخلاص التفاصيل المحلية إلى الفرع الالتفافي، بما يحقق توزيعًا وظيفيًا واضحًا للأدوار داخل النموذج.

ويختلف هذا التوجّه جذريًا عن مقاربة Swin Transformer، التي تمثّل إعادة تصميم كاملة للمحوّل البصري بدل اقتطاعه. إذ يستبدل Swin الانتباه الذاتي العالمي بانتباه محلي ضمن نوافذ متحركة، ويعتمد بنية هرمية متعددة المقاييس تحاكي الخصائص البنيوية للشبكات الالتفافية. وبذلك، يسعى Swin Transformer إلى أن يكون عمودًا فقريًا مستقلاً للرؤية الحاسوبية، قادرًا على توليد تمثيل شامل من طرف واحد، albeit بكلفة حسابية أعلى نسبيًا. في المقابل، لا يهدف المحوّل البصري المقتطع إلى استبدال الشبكات الالتفافية، بل إلى مرافقتها بوصفه مكمّلًا سياقيًا خفيفًا، يُستخدم حيث تكون العلاقات بعيدة المدى ضرورية دون فرض عبء معماري شامل.

ومن منظور منهجي، يعكس الفرق بين المحوّل المقتطع وSwin Transformer اختلافًا في الفلسفة المعمارية أكثر من كونه اختلافًا في الأداء فقط. فبينما يجسّد Swin محاولة لجعل المحوّل البصري بديلًا عامًا للشبكات الالتفافية، يمثّل الاقتطاع توجّهًا تقشّفيًا واعيًا يسعى إلى توظيف المحوّلات بقدر الحاجة فقط. وفي سياق إعادة التعرّف على الأشخاص، حيث تكون أحجام البيانات محدودة ومتطلبات النشر العملي عالية، يثبت هذا النهج الهجيني المقتطع ملاءمته بوصفه حلًا وسيطًا ناضجًا في مسار ما بعد OSNet، وممهّدًا للانتقال نحو نماذج أكثر شمولًا تتجاوز المظهر البصري وحده.



الشكل 21 بنية TE-TRANSREID

تُظهر النتائج التجريبية المعلنة في الأدبيات أن كلًّا من OSNet وTE-TransReID يحققان دقّة Rank-1 **متقاربة تقارب 95%** على مجموعة البيانات Market-1501 ضمن الإعداد القياسي، وهو ما يعكس حالة التشبّع المعماري التي بلغتها نماذج إعادة التعرّف المعتمدة على المظهر في هذا السياق. غير أنّ الفارق الجوهري بين النموذجين لا يتمثل في الأداء المطلق، بل في الكلفة الحسابية والفلسفة المعمارية؛ إذ يعتمد OSNet على بنية التفاف خفيفة متعددة المقاييس تحقق كفاءة عالية داخل إطار CNN خالص، بينما يحافظ TE-TransReID على مستوى الدقّة نفسه عبر معمارية هجينة تدمج تمثيلًا سياقيًا عالميًا باستخدام محوّل بصري مقتطع، مع تقليل ملحوظ في عدد المعاملات مقارنةً بالنماذج التحويلية الثقيلة كما تُظهر النتائج التجريبية المعلنة في الأدبيات أن كلًّا من OSNet وTE-TransReID يحققان دقّة Rank-1 متقاربة تقارب 95% على مجموعة البيانات Market-1501 ضمن الإعداد القياسي، وهو ما يعكس حالة التشبّع المعماري التي بلغتها نماذج إعادة التعرّف المعتمدة على المظهر في هذا السياق. غير أنّ الفارق الجوهري بين النموذجين لا يتمثل في الأداء المطلق، بل في الكلفة الحسابية والفلسفة المعمارية؛ إذ يعتمد OSNet على بنية التفاف خفيفة متعددة المقاييس تحقق كفاءة عالية داخل إطار CNN خالص،

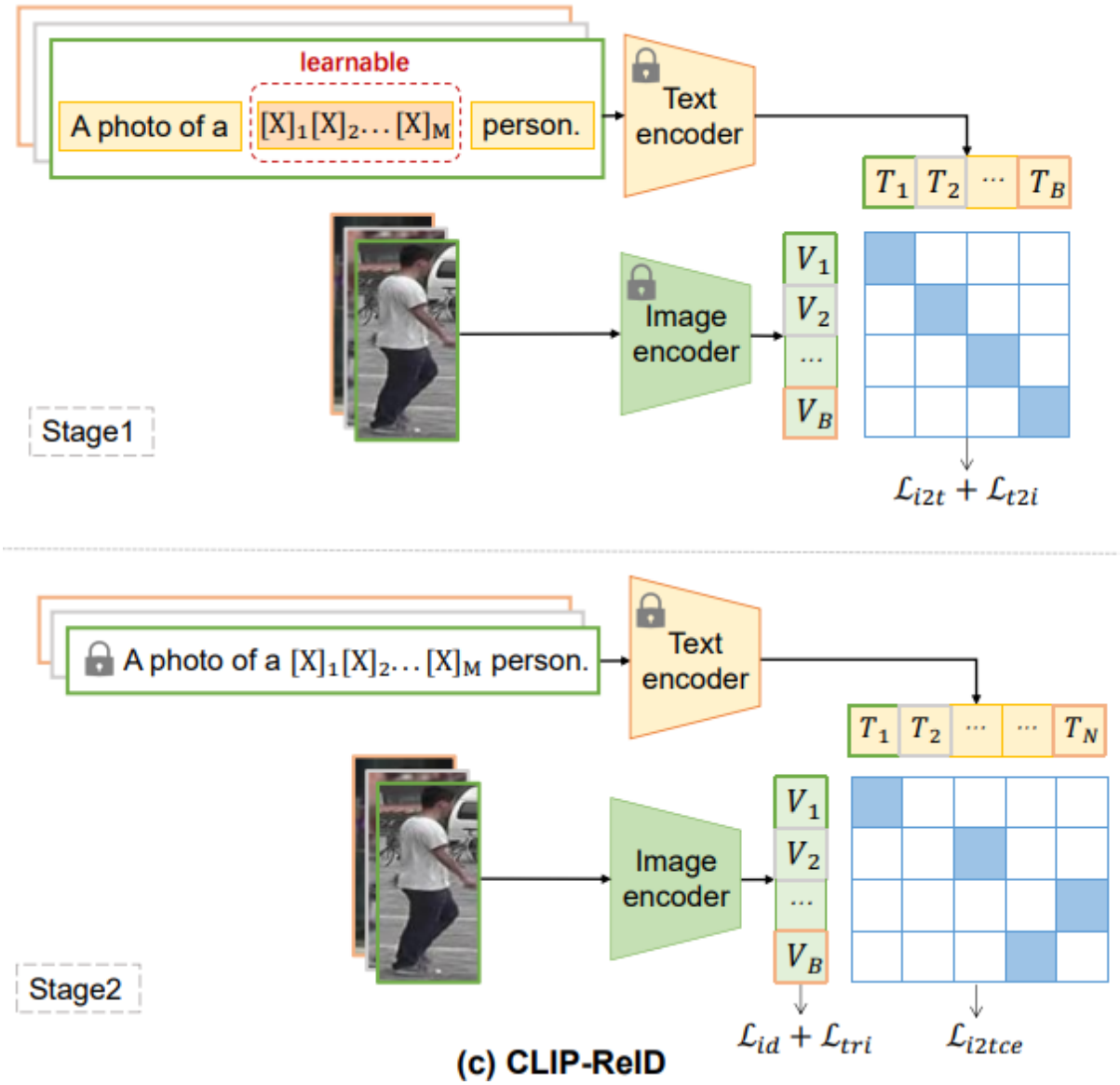
بينما يحافظ TE-TransReID على مستوى الدقّة نفسه عبر معمارية هجينة تدمج تمثيلًا سياقيًا عالميًا باستخدام محوّل بصري مقتطع، مع تقليل ملحوظ في عدد المعاملات مقارنةً بالنماذج التحويلية الثقيلة.

3.6 السمات الدلالية في مسألة إعادة التعرف

في إعادة التعرّف على الأشخاص، تشير **السمات الدلالية** إلى الخصائص التي تصف الشخص بمعناه العام وليس بمظهره اللحظي فقط، مثل الفئة العمرية، البنية الجسدية، ونمط أو نوع اللباس، وهي سمات أكثر ثباتًا وأقل تأثرًا بتغيّر الإضاءة أو زاوية التصوير أو الكاميرا مقارنة بالسمات البصرية منخفضة المستوى كالألوان والقوام. ويساعد الاعتماد على هذه السمات في تحسين التمييز بين الأشخاص المتشابهين شكليًا وتقليل تأثير التشويش البيئي. وفي هذا السياق، تسهم النماذج اللغوية–البصرية في تلخيص محتوى الصورة ضمن تمثيل دلالي مكثّف يركّز على السمات الجوهرية المرتبطة بالهوية، دون استخدام أوصاف نصية صريحة، بل عبر تمثيل عددي يوجّه الانتباه البصري ويحدّ من الاعتماد على السمات السطحية المتغيرة. ونتيجة لذلك، تتحول عملية إعادة التعرّف من مقارنة تشابه بصري مباشر إلى مقارنة تمثيل دلالي للهوية، مما يعزز متانة ودقة أنظمة Re-ID في البيئات الواقعية متعددة الكاميرات.

3.6.1. نموذج CLIP-REID [28]

يُعد نموذج CLIP (Contrastive Language–Image Pre-training) أحد أوائل النماذج التي وحّدت تمثيل الصور والنصوص ضمن فضاء دلالي مشترك عبر التعلّم التبايني على نطاق واسع. يقوم CLIP بتدريب مُشفّر بصري ومُشفّر نصي بالتوازي بحيث تُقارب تمثيلات الصورة والنص المتطابقين وتُباعد غير المتطابقين، ما يمنحه قدرة قوية على ربط المفاهيم البصرية باللغة دون الحاجة إلى إشراف خاص بكل مهمة. وقد أتاح هذا التصميم الاستفادة من المعرفة الدلالية المكتسبة من بيانات ضخمة في مهام متعددة، مثل التصنيف الصفري (Zero-shot) واسترجاع الصور، وجعل CLIP نقطة انطلاق مؤثرة لتطوير مقاربات هجينة في إعادة التعرّف تستفيد من الدلالة اللغوية إلى جانب السمات البصرية.



الشكل 22 بنية وطريقة عمل CLIP-ReID

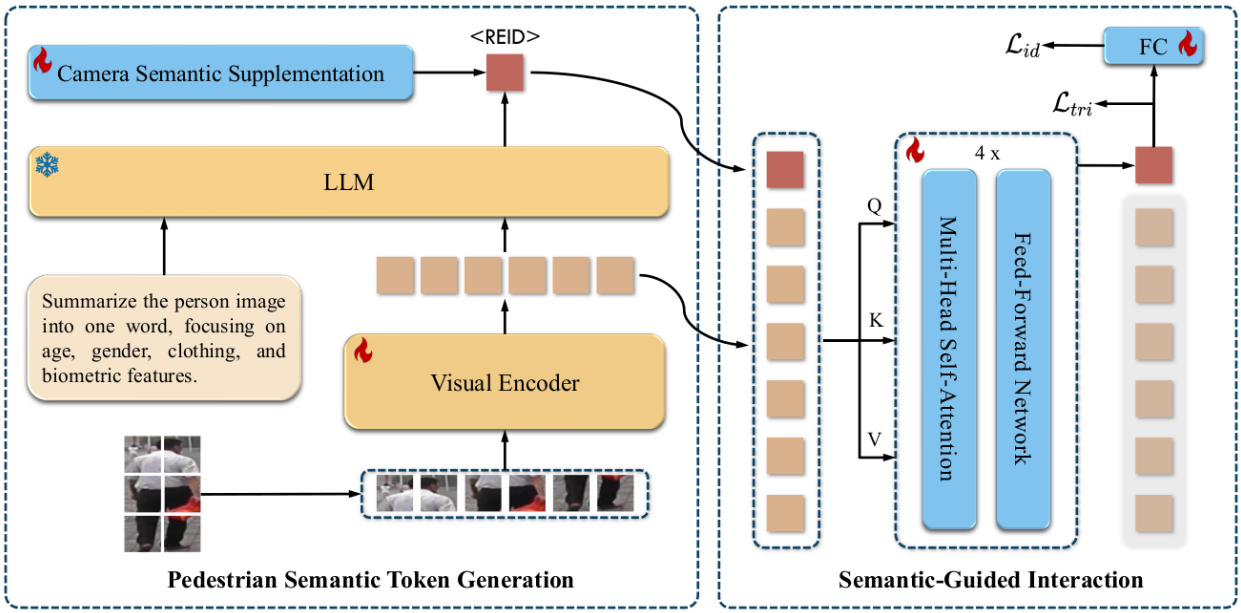
يعتمد نموذج CLIP-ReID على استثمار الفضاء الدلالي المشترك الذي تعلّمه نموذج CLIP من خلال التدريب التبايني واسع النطاق بين الصور والنصوص، وذلك لإدخال بعد دلالي إلى مهمة إعادة التعرّف على الأشخاص دون الحاجة إلى أوصاف لغوية بشرية صريحة. تتكوّن بنية النموذج من مُشفّر بصري ومُشفّر نصي مأخوذين مباشرة من CLIP، حيث يتم أثناء التدريب إنشاء قالب نصي ثابت تُدرج داخله رموز نصية قابلة للتعلّم ومخصّصة لكل هوية، تمثل معاملات داخلية لا تحمل معنى لغويًا مباشرًا. في المرحلة الأولى من التدريب، تُجمَّد مشفّرات CLIP ويُدرَّب فقط تمثيل هذه الرموز بحيث تصبح التمثيلات النصية الناتجة عنها قريبة في الفضاء المشترك من الصور التابعة للهوية نفسها وبعيدة عن صور الهويات الأخرى. بعد ذلك، وفي المرحلة

الثانية، تُستخدم هذه التمثيلات النصية المتعلَّمة كقيود دلالية لتنظيم تدريب المشفّر البصري عبر خسائر إعادة التعرّف القياسية مثل تصنيف الهوية وخسارة الثلاثيات. وعلى الرغم من إدخال اللغة في مرحلة التدريب، فإن النموذج يعتمد أثناء الاستدلال على التمثيلات البصرية فقط لإجراء المطابقة بين صور الاستعلام والمعرض، مما يجعله يحافظ على بساطة وكفاءة أنظمة Re-ID التقليدية مع الاستفادة من التنظيم الدلالي الذي توفره نماذج الرؤية—اللغة.

3.6.2. اطار عمل LVLM_ReID [29]

على الرغم من أن CLIP-ReID يُمثّل خطوة مهمة نحو إدخال الدلالة اللغوية في أنظمة إعادة التعرّف، إلا أن دوره يظل محدودًا ومشروطًا بمرحلة التدريب فقط. فالدلالة اللغوية في CLIP-ReID تُستخدم على نحو غير مباشر بوصفها قيدًا تنظيميًا للتمثيل البصري، من خلال رموز نصية خاصة بالهوية لا تحمل معنى لغويًا قابلًا للتفسير، كما يتم تجاهل المكوّن اللغوي بالكامل أثناء مرحلة الاستدلال. ونتيجة لذلك، يبقى تمثيل الهوية النهائي بصريًا بالأساس، ولا يستفيد فعليًا من قدرات النماذج اللغوية على التجريد والفهم السياقي، ولا يعبّر عن السمات الدلالية العامة التي يدركها الإنسان عند تمييز الأشخاص.

في المقابل، يعالج LVLM-ReID هذا القصور عبر دمج النموذج اللغوي–البصري الكبير كعنصر بنيوي فاعل في تمثيل الهوية نفسه، وليس كمُدرِّب خارجي. فبدل الاعتماد على رموز نصية غامضة مرتبطة بالهوية، يقوم LVLM-ReID بتوليد رمز دلالي خاص بكل صورة مستخرج مباشرة من محتواها البصري باستخدام تعليمات لغوية موجّهة، بحيث يعكس هذا الرمز السمات العامة والمستقرة للشخص، مثل نمط المظهر والبنية الجسدية. ويُدمج هذا الرمز الدلالي تفاعليًا مع السمات البصرية عبر وحدات Transformer، مما يسمح بتبادل ثنائي الاتجاه بين المعنى والدقة البصرية. والأهم من ذلك، أن هذا التمثيل الدلالي–البصري يُستخدم في كلٍ من التدريب والاستدلال، ما يحوّل عملية إعادة التعرّف من مطابقة تشابه بصري محدود إلى مطابقة هوية ذات معنى، ويمنح النظام قدرة أعلى على التعميم والثبات عبر الكاميرات والبيئات المختلفة.



الشكل 23 بنية وآالية عمل إطار العمل LVLM-REID

يعتمد هذا الإطار على استخدام تعليمات واضحة لتوجيه النموذج اللغوي الكبير المجمَّد نحو التركيز على دلالات بصرية محددة داخل صور المشاة، مما يؤدي إلى توليد رمز دلالي واحد يلخّص معلومات مظهر الشخص. بعد ذلك، يتم تصميم وحدة تفاعل فعّالة لتسهيل عملية الصقل المتبادل بين الرمز الدلالي المولَّد والرموز البصرية. وأخيرًا، يُعاد تحسين هذا الرمز المعزَّز بوصفه واصفًا مميّزًا للهوية، ويُستخدم في مهمة استرجاع الأشخاص.

المقصود هنا أن النموذج اللغوي الكبير (LLM) لا يتم تدريبه من جديد، بل يبقى مجمّد المعاملات، ويتم توجيهه فقط عبر تلقينات لغوية (Prompts) مصاغة بعناية.

هذه التعليمات تُخبر النموذج بما يجب التركيز عليه في الصورة، مثل:

- السمات العامة للمظهر
- نوع اللباس
- البنية الجسدية
- الخصائص المستقرة للشخص

أي أن اللغة تُستخدم كأداة توجيه فكري للنموذج، لا كبيانات تدريب جديدة.

بدل أن ينظر النموذج إلى الصورة كوحدات بكسل أو تفاصيل سطحية، يتم توجيهه للتركيز على الدلالات البصرية المهمة للهوية، أي *ما الذي يميّز هذا الشخص ككيان*، وليس مجرد ألوان أو حواف.

هذا يشمل مثلًا:

- النمط العام للملابس
- وجود عناصر مميّزة (حقيبة، معطف)
- الانطباع العام عن المظهر

نتيجة هذا التوجيه، يقوم النموذج اللغوي–البصري بتوليد رمز دلالي واحد (Semantic Token)، وهو:

- تمثيل عددي (ليس نصًا مقروءًا)
- يلخّص مظهر الشخص بالكامل
- يعمل كـ وصف دلالي مضغوط للهوية

يمكن اعتباره "ملخّصًا معنويًا" للشخص مستخرجًا من الصورة.

بعد توليد الرمز الدلالي، لا يتم استخدامه بشكل منفصل، بل يُمرَّر إلى وحدة تفاعل مصمَّمة خصيصًا (عادةً تعتمد على Transformer).

وظيفة هذه الوحدة هي:

- دمج الرمز الدلالي مع الرموز البصرية المستخرجة من الصورة
- السماح بتفاعل ثنائي الاتجاه:
  - الدلالة توجّه الرؤية
  - والرؤية تُثري الدلالة

يوضّح إطار LVLM-ReID كيفية توظيف التعليمات اللغوية لتوجيه نموذج لغوي–بصري مجمّد نحو استخراج تمثيل دلالي مضغوط يلخّص مظهر الشخص من الصورة. ويُدمج هذا الرمز الدلالي تفاعليًا مع السمات البصرية عبر وحدة تفاعل فعّالة، مما يسمح بصقل متبادل بين المعنى والدقة البصرية. ويُستخدم الرمز الناتج بوصفه واصفًا مميّزًا للهوية في عملية استرجاع الأشخاص، محققًا تمثيلًا أكثر ثباتًا ودلالة مقارنة بالمقاربات البصرية التقليدية.

مقارنة عامة بين CLIP-ReID و LVLM-ReID

البند	CLIP-ReID	LVLM-ReID
سنة الإصدار	2022 (AAAI 2023)	2024
نوع النموذج اللغوي	CLIP (Vision–Language Model تمثيلي)	LVLM (Large Vision–Language Model توليدي/تفسيري)
دور اللغة	تنظيم التدريب فقط (لا تُستخدم في الاستدلال)	جزء بنيوي من التمثيل (تُستخدم في التدريب والاستدلال)
طبيعة الإشارة اللغوية	رموز نصية قابلة للتعلّم خاصة بالهوية (ID-specific tokens)	تعليمات لغوية موجّهة + رمز دلالي مُولّد لكل صورة
هل النص قابل للتفسير البشري؟	لا (تمثيل عددي غامض)	جزئيًا (مستند إلى فهم لغوي/دلالي)
المشفّر البصري	ViT من CLIP	ViT داخل LVLM
المشفّر اللغوي	Text Encoder من CLIP (مجمّد)	LLM مجمّد (مع تمرير تدرّج)
مكوّنات إضافية	تعلّم على مرحلتين (Stage-1/Stage-2)	PSTG (توليد رمز دلالي) ، CSS (وعي الكاميرا) ، SGI (تفاعل دلالي–بصري)
تمثيل الهوية النهائي	Embedding بصري مُنظَّم دلاليًا	رمز دلالي–بصري مُعزَّز
كلفة الاستدلال	منخفضة	أعلى نسبيًا

قابلية التعميم الدلالي	محدودة	أعلى

الجدول 13 جدول مقارنة ين CLIP-ReID و LVLM-ReID

النموذج	مجموعات البيانات	نمط البيانات
CLIP-ReID	Market-1501، DukeMTMC-ReID، CUHK03	صور ثابتة (Image-based Re-ID)
LVLM-ReID	Market-1501، DukeMTMC-ReID، CUHK03	صور ثابتة (Image-based Re-ID)

الجدول 14 ين CLIP-ReID و LVLM-ReID من ناحية مجموعات البيانات

مجموعة البيانات	CLIP–ReID	LVLM–ReID
Market-1501	94.0	94.5
DukeMTMC-ReID	90.7	**92.2**
CUHK03	83.8	**84.6**

الجدول15 ين CLIP-ReID و LVLM-ReID على صعيد الجودة

يُظهر الجدول الأخير الخاص بنتائج Rank-1 أن كلا النموذجين يحققان أداءً مرتفعًا على مجموعات بيانات Re-ID المعتمدة على الصور، مع تفوّقٍ ملحوظ لنموذج LVLM-ReID مقارنةً بـ CLIP-ReID، ولا سيما على مجموعتي DukeMTMC-ReID و CUHK03. ويُعزى هذا التحسّن إلى قدرة LVLM-ReID على دمج المعلومات الدلالية المستخرجة لغويًا ضمن تمثيل الهوية نفسه، بدل استخدامها كقيد تدريبي غير مباشر كما في CLIP-ReID. وعلى الرغم من أن الفارق في الأداء على Market–1501 محدود، وهو ما يعكس تشبّع هذا المعيار نسبيًا، فإن التحسّن الأكثر وضوحًا في البيئات ذات التغيّر العالي بين الكاميرات يشير إلى أن التمثيل الدلالي–البصري في LVLM-ReID يوفّر متانة أفضل وقدرة أعلى على التعميم، خاصة في السيناريوهات الأكثر تحدّيًا.

في إطار LVLM-ReID، تم التعامل مع معرّف الكاميرا بوصفه عامل تحكّم دلالي يهدف إلى فصل تأثيرات التصوير عن تمثيل الهوية، وذلك عبر آلية صريحة تُسمّى Camera Semantic Supplementation (CSS). عمليًا، يُخصَّص لكل كاميرا تضمين قابل للتعلّم (Camera Embedding) يمثّل الخصائص الثابتة نسبيًا لتلك الكاميرا (زاوية الرؤية، الإضاءة، النمط اللوني)، ثم يُدمج هذا التضمين داخل خط الأنابيب البصري. تقترح الورقة صيغتين للدمج، وتُظهر أن الدمج المبكّر عبر إضافته إلى تمثيلات الرقع/المدخلات البصرية قبل استخراج السمات يعطي أفضل النتائج، لأنه يتيح للنموذج تعويض الانحرافات المرتبطة بالكاميرا منذ المراحل الأولى للاستخلاص. وبهذا يصبح الرمز الدلالي المولَّد لاحقًا (عبر PSTG) أقلّ تحيّزًا لمصدر الالتقاط، وتعمل وحدة التفاعل الدلالي–البصري (SGI) على صقل تمثيل الهوية اعتمادًا على سمات جوهرية أكثر ثباتًا. النتيجة هي تحسين التعميم عبر الكاميرات وتقليل الأخطاء الناتجة عن تحوّل المجال، دون تحميل النموذج تكلفة استدلال إضافية كبيرة.

3.7. خاتمة

يستعرض هذا الفصل دراسة مرجعية شاملة تتناول الأدبيات وأبحاث السنوات الأخيرة في المتعلقة بمسألتنا وقمنا بتقسيم الدراسة وفقاً للخاصية المميزة المعتمدة ثم وفقاً للترتيب الزمني. تؤكد هذه الدراسة على أهمية المسألة وهو ما يتضح من كمية الأبحاث المتعلقة بها وتسارعهاخاصة في السنوات الأخيرة كما أنها تشكل مرجعاً مفصلاً في موضوع خوارزميات ونماذج الذكاء الصنعي المستخدمة بهدف إعادة التعرف.

تؤكد الدراسة على الإشباع الحاصل في نماذج إعادة التعرف باستخدام الصور الخام . وعلى خصوصية البيانات التي تحتاجها نماذج المشية، كما تؤكد على كون اعتماد النماذج اللغوية كمقاربات لحل مسألة اعادة التعرف مجالاً بحثياً واعداً.

الفصل الرابع: الإطار العملي

4.1 مقدمة

انطلق هذا البحث منذ مراحله الأولى من فرضية أساسية مفادها إمكانية تحويل المقاربة المقترحة إلى نظام عملي قابل للتطبيق، حيث كانت هذه الفكرة هي الحاضرة والموجهة لسير العمل في هذا البحث، الأمر الذي فرض مجموعة من القيود المنهجية والتقنية التي أسهمت في تحديد نطاق الدراسة واتجاهاتها.

وتتمثل القيود الرئيسة للمسألة المدروسة في الاعتماد، قدر الإمكان، على بيئات الحوسبة الحافية (Edge Computing)، بما ينسجم مع متطلبات الأنظمة الواقعية ذات الموارد المحدودة. كما افترضت الدراسة بيئة تشغيل داخلية في المقام الأول، وذلك بهدف تقليل أثر العوامل الخارجية غير المتحكم بها، والتركيز على تحليل أداء النماذج ضمن ظروف تشغيل أكثر استقرارًا. إضافةً إلى ذلك، فقد تم تقييد زمن إعادة المطابقة ليكون ضمن إطار زمني محدود من رتبة الساعات، بما يتوافق مع سيناريوهات الاستخدام العملية للأنظمة الذكية في البيئات المغلقة.

إن هذه المحدودية في زمن التعرف واستخراج الأنماط ومن ثم التعرف وإعادة التعرف هي المعايير التي نستطيع بالبناء عليها اعتبار الخصائص المميزة التي قمنا باسبر امكانياتها ودراسة حالاتها في الدراستين المرجعية والعملية خصائص بيومترية تحقق شروط الخصائص البيومترية عموماً ونخص بالذكر منها الثبات (Permanence).

وفي ضوء هذه القيود، تركز التوجه البحثي على دراسة وتحليل الخصائص المميزة المختلفة المستخدمة في مسألة إعادة المطابقة، ومقارنة المقاربات المتنوعة المعتمدة في استخراج هذه الخصائص من حيث الدقة الحسابية، وكفاءة التشغيل، ومتطلبات الموارد. كما شمل البحث تحليل الحدود العملية لكل مقاربة، والسعي إلى إيجاد توازن مدروس بين الأداء والدقة من جهة، وقابلية التشغيل في البيئات المحدودة من جهة أخرى. وأخيرًا، تم التطرق إلى دراسة إمكانية التكامل بين هذه المقاربات، وتحليل أثر دمجها ضمن إطار موحد على تحسين الأداء الكلي للنظام المقترح.

4.2 استخراج المعطيات المناسبة كمدخلات لنماذج إعادة التعرف المعتمدة على المشية

كما ذكرنا سابقاً فإن النماذج المعتمدة على التقاط وتعرف أنماط التغييرات التي تميز المشية فإن مدخلات هذه النماذج هي إما تغيرات في البيان المعبر عن وضعية الجسم أو التغييرات التي تطرأ على الصورة الطلية للجسم بين إطار وأخر، لذا فإنه من الضرورة بمكان استخراج هذه المعطيات من الصور التي يتم التقاطها من كاميرات المراقبة قبل تمريرها كمدخلات لنماذج المشية

فيما يلي نستعرض النماذج والطرائق التي درسناها في استخراج الوضعيات والصور الظلية.

4.2.1. استخراج الوضعيات Pose Extraction

إن استخراج الوضعية البشرية (Pose Extraction) هو خطوة محورية في الحل الذي قمنا باقتراحه.

وفي محاولة سبر الخيارات بين الدقة والسرعة قمنا بالقيام بتحليل مقارن بين نموذجين لاستخراج الوضعيات هما HRNet، YOLOv8-Pose بهدف المفاضلة بين السرعة والجودة.

نذكر أننا لم نستطع الحصول على مجموعة بيانات تتضمن صورة لأشخاص مع الهياكل الممثلة للوضعيات بها، لذلك قمنا بالانطلاق من معرفتنا بأن نموذج HRNet هو الأكثر دقة بين النموذجين وقمنا باعتبار نتائجه على أنها هي المعطيات المرجعية أو خط الأساس للمقارنة مع نموذج Yolo8-pose قمنا بتصميم التجربة كما يلي بهدف توفير تقييم شامل وموضوعي قدر الإمكان.

في إطار تقييم الدقة، يمثل نموذج الخادم (Server Model) مرجعًا أساسيًا بدقته العالية، وتُستخدم مخرجاته كـ "حقيقة مرجعية". في المقابل، يمثل نموذج YOLOv8-Pose نموذج الحوسبة الطرفية (Edge Model)، الذي تم تحسينه خصيصًا لتحقيق السرعة.

مجموعات البيانات:

لضمان تغطية واسعة للتحديات والبيئات المختلفة في أبحاث إعادة التعرّف، أُجريت التجارب على ثلاث مجموعات بيانات قياسية وشائعة الاستخدام:

- Market-1501
- DukeMTMC-reID
- CUHK03

حجم العينة:

تم تحليل 30 صورة من كل مجموعة بيانات، ليصل إجمالي المقارنات المرئية والمترية إلى 90 مقارنة.

مقاييس التقييم:

اعتُمدت ثلاثة مقاييس رئيسية لتقييم أداء النموذجين:

- معدل اكتشاف النقاط الرئيسية (Keypoint Detection Rate): يقيس النسبة المئوية للنقاط الهيكلية (من أصل 17 نقطة) التي نجح كل نموذج في تحديدها.
- مستوى الثقة (Confidence Level): يمثل متوسط درجة الثقة التي يقدمها النموذج لدقته في تحديد النقاط الرئيسية.

الانحراف الموضعي (Positional Deviation): يقيس متوسط الخطأ بالبكسل في تحديد موضع النقاط الرئيسية لنموذج YOLOv8-Pose، وذلك مقارنة بالمواقع المرجعية المستخرجة من HRNet33.

كانت نتائج التجربة كما يلي:

استطاع كلا النموذجين من اكتشاف أو استيفاء كافة النقاط الـ 17 في 100% من العينات، ومنه نستنتج أن كلا النموذجين فعالان في تحديد جميع النقاط الرئيسية المطلوبة. الفارق لا يكمن في القدرة على الاكتشاف، بل في دقة هذا الاكتشاف.

مستويات الثقة

يوضح مستوى الثقة مدى "يقين" النموذج من صحة توقعاته.

HRNet: أظهر موثوقية عالية بمتوسط ثقة بلغ 0.901 (أي ~90%).

YOLOv8-Pose: أظهر موثوقية متوسطة بمتوسط ثقة بلغ 0.800 (أي ~80%).

الاستنتاج: يمثل الانتقال إلى نموذج YOLOv8-Pose انخفاضًا بنسبة 10% تقريبًا في متوسط الثقة مقارنةً بـ HRNet. هذا الانخفاض هو الثمن المباشر للحصول على سرعة استنتاج أعلى.

الدقة الموضعية والانحراف

يُعد هذا المقياس هو الأكثر أهمية لتقييم جودة السمات الهيكلية.

المتوسط العام: بلغ متوسط الانحراف الموضعي لنموذج YOLOv8-Pose (مقارنةً بـ HRNet) 15.38 بكسل عبر جميع مجموعات البيانات.

تم تحليل الأداء أيضًا بشكل منفصل لكل مجموعة بيانات، مما كشف عن تباين يعتمد على مدى "تحدي" مجموعة البيانات (مثل الازدحام، الإضاءة، وزوايا الكاميرا):

Market-1501 (الحالة الأفضل): 9.97 بكسل انحراف.

DukeMTMC-reID (الحالة المتوسطة): 15.31 بكسل انحراف.

CUHK03 (الحالة الأكثر تحديًا): 20.87 بكسل انحراف.

الاستنتاج: الانحراف ليس ثابتًا، بل يزداد مع صعوبة مجموعة البيانات. تُظهر مجموعة بيانات CUHK03 أن الانحراف يمكن أن يتجاوز 20 بكسل، وهو ما قد يكون له تأثير ملموس على السمات الهندسية الدقيقة.

مناقشة النتائج وتطبيقاتها

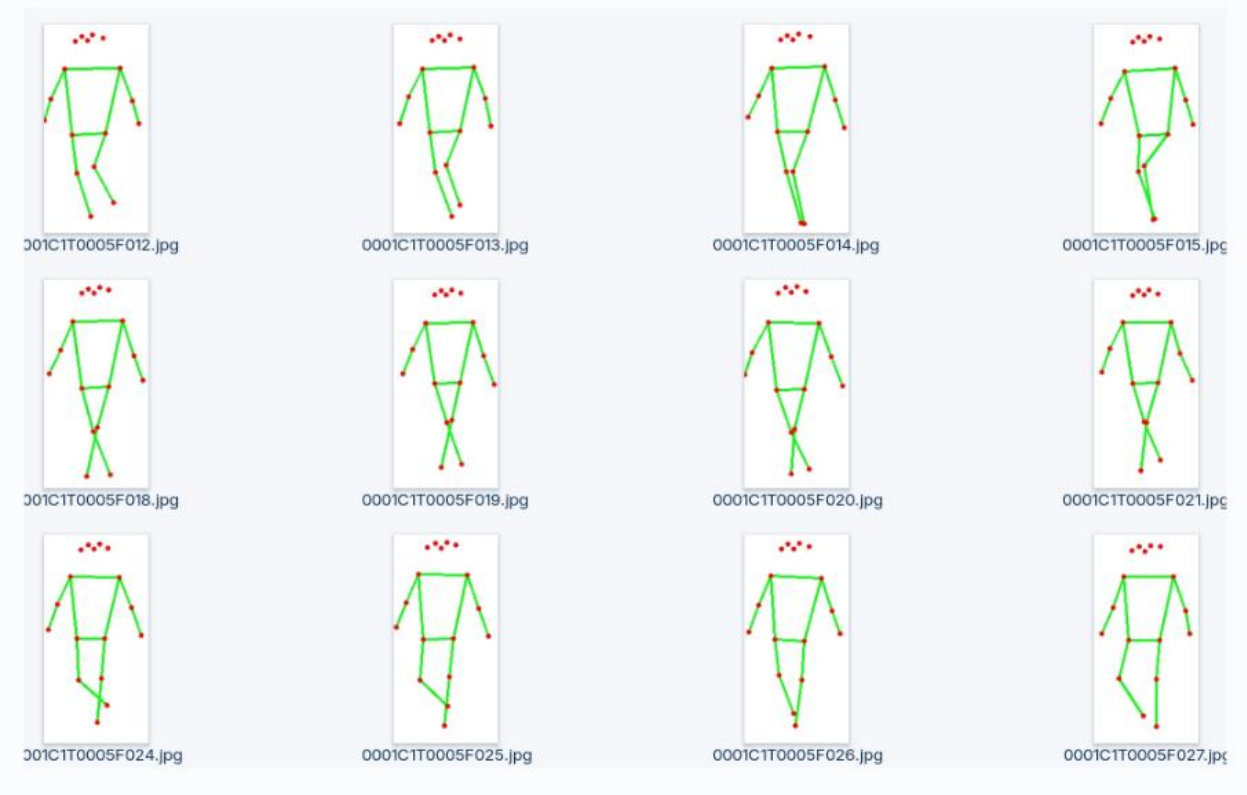
توفر هذه النتائج الكمية أساسًا لاتخاذ قرارات هندسية مستنيرة بخصوص بنية نظام إعادة التعرّف (Re-ID) المعتمد على السمات الهيكلية.

المفاضلة في اختيار النموذج

أكدت النتائج وجود مفاضلة واضحة:

يُنصح باختيار HRNet عندما تكون الدقة هي الأولوية القصوى، ولا توجد قيود زمنية صارمة (مثل المعالجة الدفعية (Batch processing) على الخادم).

يُنصح باختيار YOLOv8-Pose عندما تكون المعالجة في الوقت الفعلي أمرًا ضروريًا (مثل التطبيقات الطرفية)، ويمكن قبول خسارة معينة في الدقة (انخفاض 10% في الثقة وانحراف ~15 بكسل).



التأثير المتوقع على تدريب نماذج إعادة التعرّف

إن الانحراف الموضعي البالغ 15.38 بكسل ليس ضئيلاً وقد يؤثر سلبًا على جودة السمات المستخرجة بواسطة النماذج المعتمدة على الوضعية، التي تعتمد على العلاقات المكانية الدقيقة بين المفاصل.

للتخفيف من هذا الأثر عند استخدام YOLOv8-Pose كمصدر للبيانات، يُقترح بشدة تطبيق تقنيات تعزيز البيانات (Data Augmentation) أثناء تدريب النموذج المعتمد على الوضعية. يمكن أن يشمل ذلك إضافة "ضجيج" (noise) أو إزاحات طفيفة إلى مواقع النقاط الرئيسية بشكل مصطنع، مما يجعل النموذج المدرب أكثر "مرونة" (robust) وقدرة على التعامل مع هذه الاختلافات في بيانات الإدخال.

سيناريوهات النشر (Deployment)

بناءً على النتائج، يمكن اقتراح البنى التالية:

النشر على الخادم (Server-side): استخدام HRNet لاستخراج السمات بأعلى جودة.

النشر على الحافة (Edge-side): استخدام YOLOv8-Pose للمعالجة الفورية في الموقع.

النشر الهجين (Hybrid): استخدام YOLOv8-Pose لإجراء اكتشاف أولي سريع، يليه تنقيح (refinement) بواسطة HRNet للحالات التي تتطلب دقة عالية.

وهنا نحتاج إلى مرشح ما يقرر مدى جودة الوضعية المستخرجة ومدى كفاية هذه الجودة وهو ما يتطلب تجارب أكثر لتحديد العتبة المناسبة.

توفر هذه المقاييس (10% انخفاض ثقة، 15.38 بكسل انحراف) البيانات اللازمة لتقييم ما إذا كانت النماذج المعتمدة على الوضعية يمكنها الحفاظ على أدائها في ظل استخدام بيانات هيكلية أقل دقة، وتوجيه الخطوات التالية نحو تدريب هذه النماذج لتكون مرنة وقوية ضد هذه الانحرافات.

4.2.2. استخراج الصور الظلية

مقدمة

أجرينا تحليلاً لأداء ست طرائق متقدمة لتجزئة الصور، مطبقة على صور المشاة، عبر أربع تركيبات معالجة مختلفة. يركز التقييم على مقاييس الأداء (السرعة) والجودة (الدقة)، بالإضافة إلى فعالية خط أنابيب المعالجة.

التركيبات الأربع المقصودة هي:

- استخراج مباشر للصورة الظلية
- معالجة سبقية ثم استخراج
- استخراج ثم معالجة لحقية
- تطبيق معالجة سبقية ومعالجة لحقية (both_processig)

الطرائق الستة هي:

- GrabCut

تعتمد خوارزمية GrabCut على نماذج ماركوف العشوائية (MRF) ونمذجة الألوان باستخدام خليط من توزيعات غاوسية (GMM) لفصل المقدمة عن الخلفية. تبدأ الخوارزمية بتحديد تقريبي للعنصر الهدف (عادةً عبر مستطيل ابتدائي)، ثم تُحسّن حدود الفصل تكراريًا من خلال تقليل دالة طاقة تجمع بين اتساق اللون وسلاسة الحواف، ما يجعلها مناسبة للسيناريوهات التفاعلية ذات الموارد المحدودة، لكنها غير ملائمة للتشغيل الآني واسع النطاق.

- YOLOv8-seg

تمثل YOLOv8-seg امتدادًا لنماذج YOLO أحادية المرحلة، حيث تجمع بين الكشف والتقسيم الدلالي/اللحظي ضمن إطار موحد. تعتمد الطريقة على شبكة عصبية تلافيفية عميقة تقوم بتوقع الصناديق المحيطة وأقنعة التقسيم مباشرة في تمريرة واحدة، مما يحقق توازنًا فعالًا بين الدقة وسرعة المعالجة، ويجعلها مناسبة لتطبيقات الزمن الحقيقي والبيئات الحافية.

- Mask R-CNN

تعتمد Mask R-CNN على بنية ثنائية المرحلة، حيث تقوم المرحلة الأولى باقتراح مناطق اهتمام (Region Proposals)، تليها مرحلة تصنيف هذه المناطق واستخراج أقنعة دقيقة لكل كائن على حدة. تتميز هذه الطريقة

بدقة عالية في تحديد الحدود الهندسية للأجسام، إلا أن كلفتها الحسابية المرتفعة تجعل استخدامها أقل ملاءمة للتطبيقات ذات القيود الزمنية أو الحاسوبية الصارمة.

- SAM

يعتمد نموذج SAM على بنية قائمة على المحوّلات (Transformers)، ويهدف إلى توفير تقسيم عام غير مخصص لمجال معين. يتميز بقدرته على توليد أقنعة دقيقة استجابةً لمحفزات مختلفة (نقاط، مربعات، أو أقنعة أولية)، ما يمنحه مرونة عالية وقابلية تعميم قوية. إلا أن هذه القوة تأتي على حساب المتطلبات الحسابية المرتفعة، خصوصًا في غياب تسريع عتادي مناسب.

- U^2-Net

يرتكز U^2-Net على مفهوم البنى المتداخلة (Nested U-Structures)، حيث يتم تضمين وحدات U-Net صغيرة داخل بنية U-Net أكبر، ما يسمح بالتقاط التفاصيل الدقيقة والسياق العالمي في آن واحد. تُستخدم هذه الطريقة بكفاءة في مهام فصل المقدمة، خاصة في الصور ذات الخلفيات المعقدة، مع كلفة حسابية أقل نسبيًا مقارنة بالنماذج الضخمة المعتمدة على المحوّلات.

- SAM2

يمثل SAM2 تطويرًا معماريًا لنموذج SAM، مع تحسينات تستهدف الكفاءة الحسابية وقابلية التوسع، خصوصًا في سيناريوهات الفيديو أو التسلسلات الزمنية. يركز النموذج على تعزيز الاستقرار الزمني للأقنعة وتحسين سرعة الاستجابة، ما يجعله أكثر ملاءمة للتطبيقات العملية مقارنة بالإصدار الأصلي، مع الحفاظ على قدرات التعميم العالية.

النتائج الرئيسية

- التركيبة الأفضل أداءً: "both_processing" (المعالجة الأولية واللاحقة).
- الطريقة الأسرع: YOLOv8-seg (تحقق أسرع زمن بمعالجة وجودة مقبولة).
- الطريقة الأكثر دقة (جودة): Mask R-CNN (تحقق أعلى دقة في اكتشاف المشاة).
- الطريقة الأكثر توازناً: U^2-Net (تُظهر جودة ثابتة عبر جميع تركيبات المعالجة).

تحليل التركيبة الأفضل أداءً (both_processing)

تُظهر تركيبة "both_processing" تفوقًا في التوازن بين السرعة والجودة، حيث تسجل أسرع متوسط زمن معالجة (1.02 ثانية) وتحافظ على مستوى جودة عالٍ. كما تحقق أعلى متوسط تغطية (26.8%)، مما يدل على اكتشاف أفضل للكائن

الأمامي، وتحسنًا في جودة الشكل من خلال درجات أعلى في مقاييس التراص والصلابة، مع إظهار نتائج مستقرة عبر جميع الطرق المطبقة.

تحليل الأداء (السرعة)

الطريقة	متوسط الزمن (ثانية)
GrabCut	0.09
YOLOv8-seg	0.20
Mask R-CNN	0.62
SAM	1.28
U^2-Net	2.20
SAM2	2.76

الجدول 16 مقارنة أداء طرائق استخراج الصور الظلية على صعيد زمن التنفيذ

استنتاجات تحليل الأداء:

- YOLOv8-seg: الأداء الأسرع إجمالاً، ومُحسن لتطبيقات الزمن الحقيقي.
- GrabCut: سرعة معالجة عالية على وحدة المعالجة المركزية، ومناسبة للسيناريوهات البسيطة.
- طرق التعلم العميق (SAM2 و U^2-Net): زمن معالجة أبطأ، ولكن جودة أعلى.

تحليل الجودة

لقياس جودة النتائج من النماذج المختبرة واتساقاً مع المسألة وهدفنا في تحديد أفضل مستخرج صورة ظلية من صور المشاة حصراً اعتمدنا المقاييس التالية.

مقاييس الجودة

يعتمد تقييم الجودة على مقاييس هندسية وإحصائية:

1. التغطية (Coverage): النسبة المئوية لبكسلات الجسم الأمامي. الأهمية للمشاة: اكتشاف أفضل للظل الكامل.
2. التراص (Compactness): يقيس مدى دائرية/تراص الشكل المجزأ. الأهمية للمشاة: ظلال أكثر اكتمالًا وأقل تجزئة، ومهم لاكتشاف الأوضاع كاملة الجسم.
3. الصلابة (Solidity): نسبة مساحة الهيكل المحدب التي يملؤها الشكل. الأهمية للمشاة: وجود ثقوب أقل في الظل، وحاسم لاكتشاف الأطراف المفقودة.
4. حدة الحواف (Edge Sharpness): مقدار التدرج عند حدود التجزئة. الأهمية للمشاة: تجزئة أكثر دقة وفصل نظيف بين الشخص والخلفية.
5. نسبة العرض إلى الارتفاع (Aspect Ratio): نسبة العرض إلى الارتفاع للمستطيل المحيط. الأهمية للمشاة: الكشف عن الأشكال المبتورة أو الممدودة.
6. الانحراف المركزي (Eccentricity): مقدار الانحراف المركزي للقطع الناقص. الأهمية للمشاة: الإشارة إلى أوضاع ممدودة أو متراصة.

أهمية هذه المقاييس لصور المشاة:

- الاكتشاف الكامل: التغطية تضمن اكتشاف الشخص بالكامل.
- سلامة الشكل: التراص والصلابة يمنعان النتائج المجزأة.
- دقة الحدود: حدة الحواف تضمن فصلًا نظيفًا.
- تحليل الوضعية: نسبة العرض إلى الارتفاع والانحراف المركزي تساعدان في فهم الوضعية.
- معالجة الانسداد: الصلابة المنخفضة تشير إلى أجزاء مفقودة أو انسدادات.

شرح طرق التجزئة

1. **GrabCut (OpenCV)**:
   - الخوارزمية: تجزئة تفاعلية تعتمد على "تقطيع الرسم البياني".
   - نقاط القوة: سريع للغاية (0.09 ثانية)، لا يتطلب GPU، جيد للخلفيات البسيطة.

- نقاط الضعف: تغطية وصلابة منخفضة، صعوبة مع الخلفيات المعقدة.
- الأفضل لـ: الكائنات البسيطة، بيئات CPU فقط، تطبيقات الزمن الحقيقي.

2. **U²-Net (rembg)**:
   - الخوارزمية: إزالة خلفية بالتعلم العميق بشبكة U-shaped.
   - نقاط القوة: تغطية وصلابة ممتازة، أداء ثابت، جيد للصور الشخصية.
   - نقاط الضعف: معالجة أبطأ (2.20 ثانية)، قد يبالغ في التجزئة.
   - الأفضل لـ: تصوير البورتريه، إزالة الخلفية، الجودة الثابتة.
3. **YOLOv8-seg (Ultralytics)**:
   - الخوارزمية: تجزئة كائنات في الزمن الحقيقي باستخدام بنية YOLO.
   - نقاط القوة: أسرع طريقة تعلم عميق (0.20 ثانية)، تغطية عالية، جيد للكائنات المتعددة، قادر على العمل في الزمن الحقيقي.
   - نقاط الضعف: صلابة أقل، قد يفقد التفاصيل الدقيقة.
   - الأفضل لـ: تطبيقات الزمن الحقيقي، اكتشاف الكائنات المتعددة، المهام التي تكون فيها السرعة حاسمة.
4. **SAM (Segment Anything Model)**:
   - الخوارزمية: نموذج تأسيسي للتجزئة القابلة للتوجيه.
   - نقاط القوة: دقة عالية عند توجيهه بشكل صحيح، بنية حديثة، جيد للمشاهد المعقدة.
   - نقاط الضعف: تغطية منخفضة بدون توجيهات مناسبة (3.9%)، معالجة أبطأ (1.28 ثانية)، يتطلب ضبطًا دقيقًا.
   - الأفضل لـ: المهام عالية الدقة، المشاهد المعقدة، التطبيقات البحثية.
5. **Mask R-CNN (Detectron2)**:
   - الخوارزمية: تجزئة كائنات ثنائية المراحل مع مقترحات مناطق.

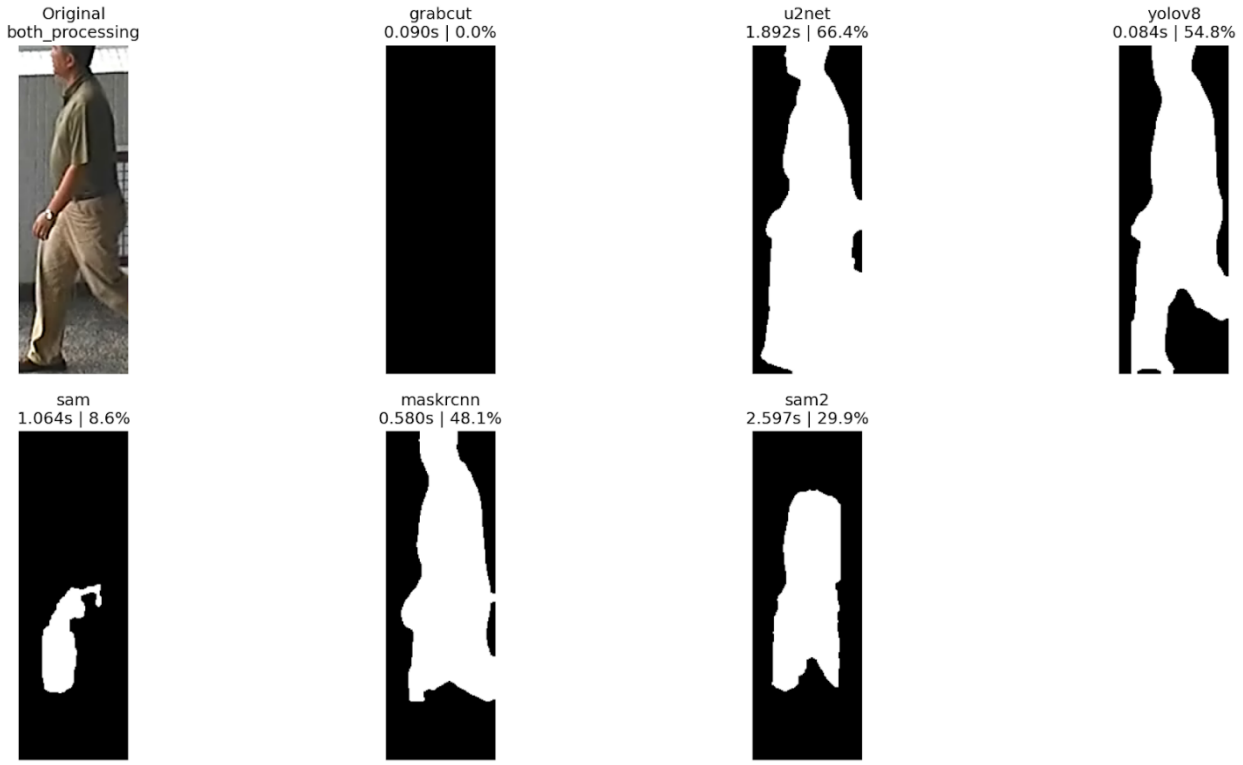
- نقاط القوة: أعلى درجات الجودة، ممتاز لاكتشاف المشاة، درجات ثقة عالية، حدود دقيقة.
- نقاط الضعف: معالجة أبطأ (0.62 ثانية)، يستهلك ذاكرة GPU بكثافة، متطلبات إعداد معقدة.
- الأفضل لـ: التطبيقات عالية الدقة، اكتشاف المشاة، الأبحاث.

6. **SAM2 (Segment Anything Model v2)**:

- الخوارزمية: نموذج تأسيسي من الجيل التالي لتجزئة الفيديو.
- نقاط القوة: جودة حديثة، صلابة جيدة، بنية متقدمة.
- نقاط الضعف: أبطأ معالجة (2.76 ثانية)، تغطية منخفضة (14.2%)، متطلبات حسابية عالية.
- الأفضل لـ: التطبيقات البحثية، متطلبات الجودة الأعلى، تجزئة الفيديو.

التصنيف	الطريقة	متوسط التغطية (%)	متوسط التراص	متوسط الصلابة	درجة الجودة (من 10)
1	Mask R-CNN	41.4	0.33	0.81	9.2
2	U^2-Net	47.6	0.34	0.8	8.8
3	YOLOv8-seg	50.4	0.27	0.73	8.5
4	SAM2	14.2	0.36	0.85	7.8
5	SAM	3.9	0.33	0.73	6.5
6	GrabCut	2.6	0.29	0.48	4.2

الجدول17 مقارنة أداء طرائق استخراج الصور الظلية على صعيد مقاييس الجودة



الشكل 24 إحدى العينات ونتائج استخراج الصورة الطلية منها بالطرائق المختبرة - تظهر بوضوح عدم جودة النتائج

كما يتضح جلياً من الشكل أعلاه ،ن صور مجموعات البيانات ليست بجودة كافية تسمح باستخراج صور ظلية جيدة دون تدخل بشري ضروري.

استنتاجات تحليل الجودة:

- Mask R-CNN: الأنسب لمهام الاكتشاف الدقيق للمشاة بدرجات ثقة عالية.
- U^2-Net: أداء ممتاز للصور الشخصية وجودة ثابتة عبر جميع تركيبات المعالجة.
- YOLOv8-seg: توازن مقبول بين السرعة والجودة لتطبيقات الزمن الحقيقي.

التوصيات

- **تطبيقات الزمن الحقيقي (أولوية للسرعة):** YOLOv8-seg مع preprocessing_only، GrabCut مع postprocessing_only، U^2-Net مع preprocessing_only.
- **تطبيقات الجودة العالية (أولوية للجودة):** Mask R-CNN مع both_processing، U^2-Net مع both_processing، SAM2 مع both_processing.

- **التطبيقات المتوازنة (سرعة + جودة):** YOLOv8-seg مع both_processing، U^2-Net مع both_processing، Mask R-CNN مع preprocessing_only.
- **البيئات محدودة الموارد:** GrabCut مع (CPU-only) postprocessing_only، U^2-Net مع no_processing (GPU متوسط).

تأثير خط أنابيب المعالجة

يشير خط أنابيب المعالجة إلى سلسلة العمليات المطبقة قبل وبعد تشغيل خوارزمية التجزئة الرئيسية.

- المعالجة الأولية (Preprocessing): عمليات تُطبق على الصورة الأصلية قبل إدخالها للنموذج (تحسين التباين، شحذ الحواف).

  الفوائد: تحسين الجودة، الاتساق.

- المعالجة اللاحقة (Postprocessing): عمليات تُطبق على قناع التجزئة الناتج (عمليات مورفولوجية، تنعيم).

  الفوائد: صقل الشكل، تقليل الضوضاء، إزالة التشوهات.

- خط الأنابيب المدمج (Combined Pipeline):

  الفوائد: الأداء الأمثل، نتائج قوية، الجاهزية للإنتاج.

ملاحظات فنية حول التنفيذ

- صيغة حساب درجة الجودة: Σ (قيمة_المقياس × الوزن × عامل_الاتجاه)
- أوزان المقاييس: التغطية (25%)، التراص (20%)، الصلابة (20%)، حدة الحواف (15%)، نسبة العرض إلى الارتفاع (10%)، الانحراف المركزي (10%).

الخلاصة

يوضح التحليل أن تركيبة "both_processing" مع YOLOv8-seg توفر أفضل توازن بين السرعة والجودة لمهام تجزئة المشاة. يعمل خط أنابيب المعالجة الأولية واللاحقة على تحسين الأداء والجودة بشكل كبير، مما يجعله ضروريًا للتطبيقات الإنتاجية.

النقاط الرئيسية:

- تحسين خط أنابيب المعالجة حاسم للحصول على أفضل النتائج.

- YOLOv8-seg يقدم أفضل توازن بين السرعة والجودة.
- Mask R-CNN يوفر أعلى دقة لاكتشاف المشاة.
- مقاييس الجودة مصممة خصيصًا لتحليل ظلال المشاة.
- يجب أن يعتمد اختيار الطريقة على متطلبات التطبيق المحددة.

تم إعداد هذا التقرير بناءً على تقييم شامل لـ 192 تجربة تجزئة عبر 8 صور اختبار باستخدام 6 طرق مختلفة و4 تركيبات معالجة.

4.3. إشكالية توافر مجموعات البيانات الشاملة لتقييم نماذج إعادة التعرّف المعتمدة على المشية

تُظهر مراجعة مجموعات البيانات المعيارية المستخدمة في الأبحاث المتعلقة بإعادة التعرّف على الأشخاص وجود تفاوت واضح في درجة التوافر وإمكانية الوصول تبعًا لطبيعة التمثيل المعتمد. فقد تبيّن أن مجموعات البيانات المعتمدة على الصور الثابتة (Image-based Re-ID datasets) والمتداولة على نطاق واسع في الأدبيات المرجعية متاحة بشكل مجاني وسهل الوصول، الأمر الذي أتاح استخدامها دون قيود تُذكر في هذا العمل. وبالمثل، وعلى الرغم من أن مجموعات البيانات الفيديوية تُعد أقل انتشارًا نسبيًا، فقد أمكن الحصول على عدد منها بما يكفي لدعم التجارب المرتبطة بالنماذج المعتمدة على التسلسل الزمني للمشاهد.

في المقابل، برزت تحديات جوهرية فيما يتعلق بمجموعات البيانات المخصصة لنمذجة المشية (Gait-oriented datasets). إذ اقتصر ما أمكن الوصول إليه فعليًا على مجموعة CASIA-B، وبشكل جزئي فقط، حيث أُتيحت الصور الظلية (Silhouettes) دون الوصول إلى الصور الأصلية بصيغة RGB، وذلك لأسباب تتعلق بقوانين حماية الخصوصية والبيانات الشخصية. كما توجد نسخة لها تشمل بيانات الوضعيات (Pose-based representations) مشتقة من هذه الصور، إلا أنها غير رسمية ولا تخضع لإطار توثيقي أو معياري واضح. أما بقية مجموعات بيانات المشية المعروفة في الأدبيات، فهي تتطلب في جميع الحالات مراسلات مباشرة مع الجهات المالكة أو الباحثين القائمين عليها، إضافة إلى توقيع اتفاقيات استخدام خاصة، الأمر الذي حال دون إمكانية الحصول عليها ضمن الإطار الزمني والإجرائي لهذا البحث.

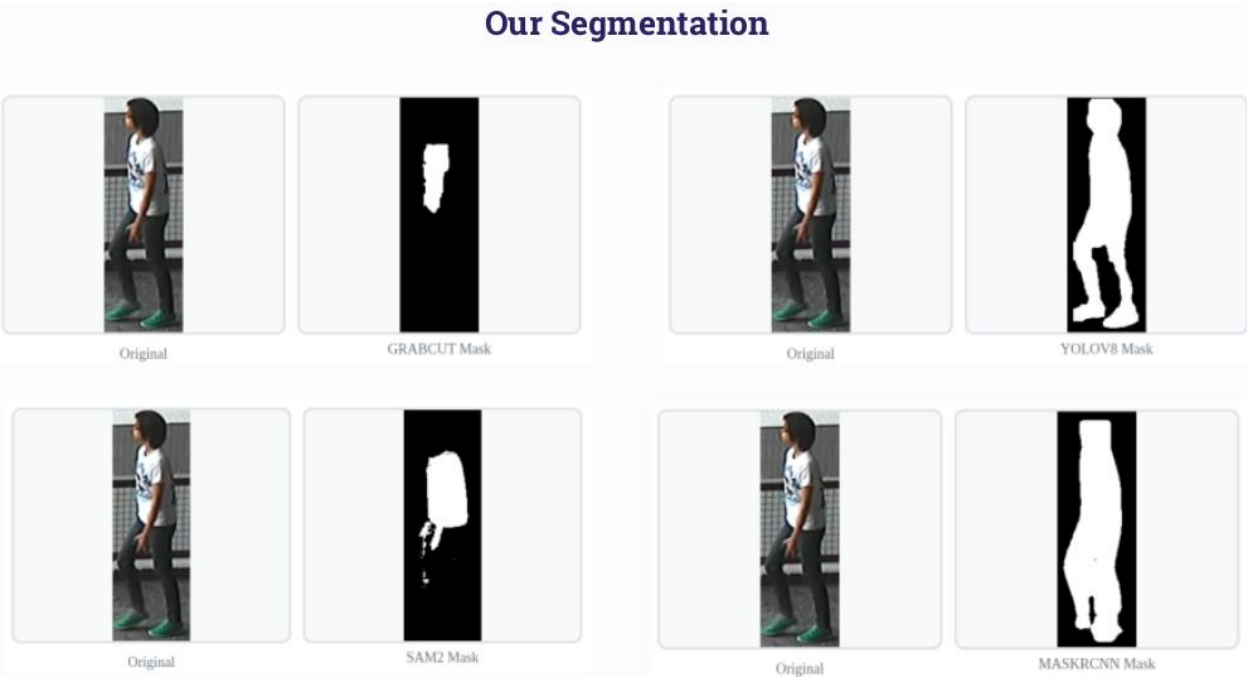
نتيجة لذلك، برزت إشكالية منهجية تتمثل في غياب مجموعة بيانات شاملة تجمع، ضمن إطار واحد متسق، بين مختلف التمثيلات اللازمة لتقييم النماذج متعددة المسارات، بما في ذلك التمثيل الحركي القائم على الصور الظلية، والتمثيل الهندسي القائم على الوضعيات، والتمثيل الشكلي المرتبط بالمظهر الخارجي، إضافة إلى السمات الدلالية المستخلصة من الوصف النصي

أو النماذج اللغوية البصرية. ويُعد هذا الغياب عائقًا مباشرًا أمام إمكانية إجراء تقييم موحّد وعادل للنماذج التي تعتمد على دمج هذه التمثيلات في آن واحد.

وعليه، فإن مجموعات البيانات التي أمكن الحصول عليها في هذا العمل تُعد مناسبة لتدريب وتقييم النماذج المعتمدة على المظهر والسمات الدلالية، لكنها لا تستوفي المتطلبات البنيوية للنماذج المعتمدة على المشية، ولا سيما تلك التي تفترض توافر معلومات زمنية وحركية متسقة عبر تسلسلات متعددة. وبناءً على هذه القيود، تم اعتماد استراتيجية تقييم تفاضلية، حيث جرى اختبار كل فئة من النماذج ضمن مجموعات البيانات التي تتوافق مع افتراضاتها التمثيلية، مع الأخذ بعين الاعتبار حدود المقارنة بين هذه الفئات عند تحليل النتائج ومناقشتها.

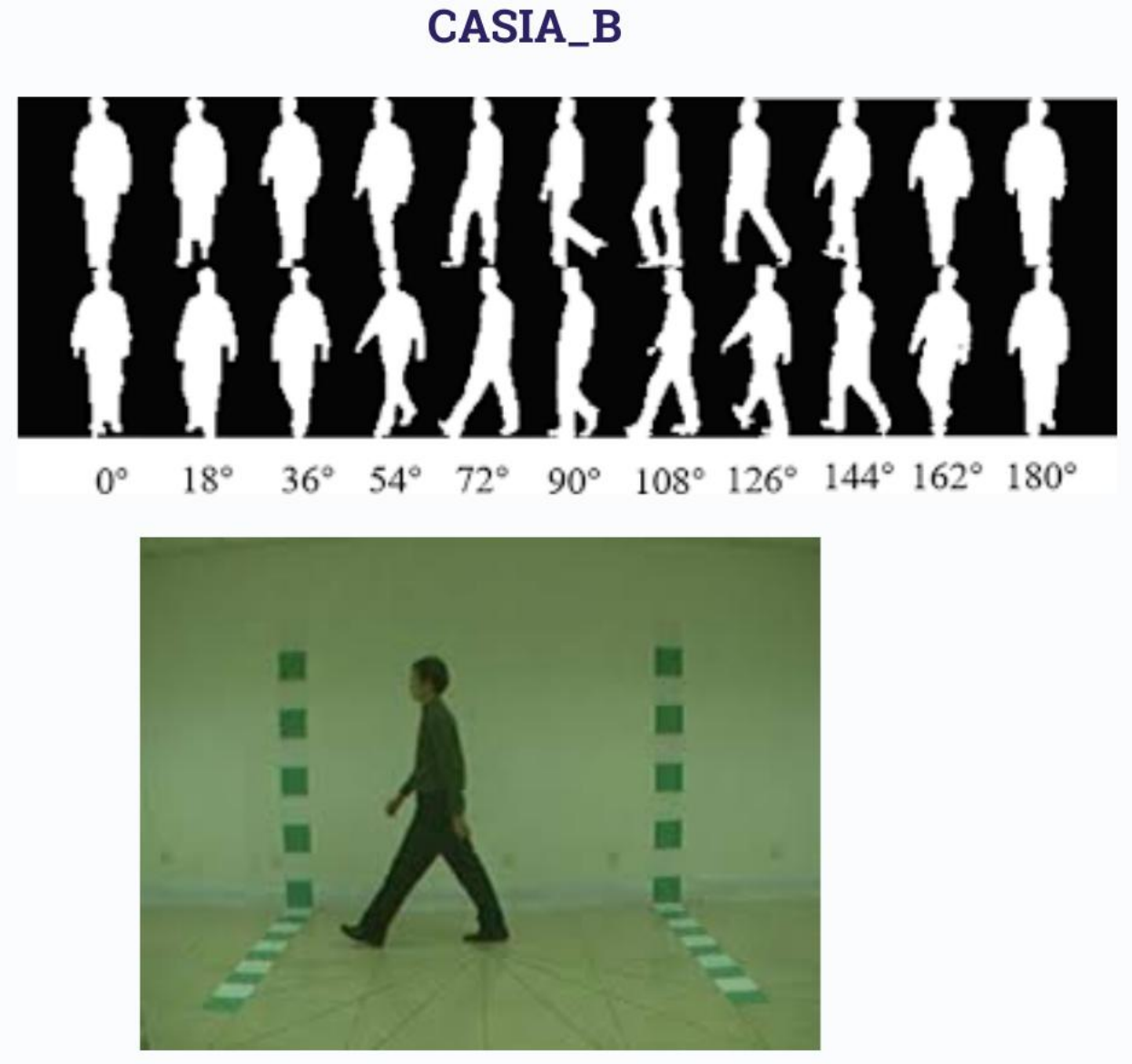
إضافةً إلى القيود المرتبطة بتوافر مجموعات البيانات، واجه هذا العمل تحديًا نوعيًا يتمثل في جودة الصور الظلية والوضعيات المستخرجة من مجموعات البيانات المتاحة. إذ أظهرت التجارب الأولية أن الصور الظلية التي أمكن توليدها اعتمادًا على تقنيات الفصل والتجزئة المطبّقة على مجموعات البيانات الصورية والفيديوية العامة كانت في معظم الحالات دون المستوى المطلوب، ولا تضاهي من حيث النقاء والدقة البنيوية الصور الظلية المتوفرة ضمن مجموعة CASIA-B. وقد تجلّت هذه الفجوة النوعية في مظاهر متعددة، من بينها التشوهات الحادة في حدود الجسم، والانقطاعات الزمنية بين الإطارات، والضوضاء الخلفية المتبقية، فضلًا عن فقدان الاتساق الحركي عبر التسلسل الزمني.

يوضح الشكل التالي الصور الظلية المستخرجة من إحدى الصور المتوفرة لدينا



الشكل 25 إحدى العينات والصور الظلية المستخرجة منها

بالمقارنة مع الصور التي توفرها Casia-b h التي تبدو وكأنها مقصوصة بعناية عند مقارنتها بنتائجنا ويوضح الشكل كذلك البيئة وزاوية التصوير الممتازة التي تم التقاط هذه الصور وفقها.



الشكل 26 CASIA - B و طريقة التقاط صورها والظلال المستخرجة من عدة زوايا

إن الانطلاق من مثل هذه الصور الظلية منخفضة الجودة لتدريب نماذج تعتمد في جوهرها على تحليل تغير المشية لا يُعد خيارًا مشجعًا من الناحية المنهجية. فضعف جودة المدخلات ينعكس مباشرة على استقرار عملية التدريب، ويؤدي إلى نتائج يصعب

تفسيرها علميًا، حيث يصبح من غير الممكن خلال دورات التطوير والتدريب المتكرر التمييز بشكل موثوق بين ما إذا كانت محدودية الأداء ناتجة عن رداءة البيانات، أو عن خصائص المعمارية المستخدمة، أو عن إعدادات التدريب مثل عدد التكرارات ومعاملات الضبط. ويؤدي هذا التداخل إلى تقويض صلاحية الاستنتاجات التجريبية، ويحد من قيمة أي تحسينات ظاهرية قد يتم تحقيقها.

بناءً على ما سبق، وبغية الحفاظ على الصرامة المنهجية وتفادي استخلاص نتائج قد تكون مضللة أو غير قابلة للتعميم، اقتصر هذا العمل في مسألة المشية على الدراسة المرجعية والتحليل النقدي للأعمال المنشورة، دون اقتراح نموذج جديد يعتمد على الصور الظلية المستخرجة محليًا. وقد أُخذ هذا القرار بوصفه خيارًا بحثيًا واعيًا يوازن بين الطموح التطبيقي ومتطلبات الموثوقية العلمية، مع الإشارة إلى أن تطوير نماذج مشية فعّالة يظل مشروطًا بتوافر مجموعات بيانات عالية الجودة ومصممة أصلًا لهذا الغرض.

4.4. نموذج القناة الحركية (Kinematic Stream)

قمنا بالاعتماد على مجموعة البيانات غير الرسمية التي تشمل التمثيلات البيانية لوضعيات صور مجموعة بيانات Casiab-B واقترحنا تماشياً مع ما درسناه في الدراسة المرجعية وكذلك بالاستئناس بمعمارية نموذج DeepGaitV2 الخاص بالصور الظلية باقتراح معمارية لنموذج يقوم بالتقاط الحركة ممثلة بتتالي بيانات الوضعيات وترميزها ضمن أشعة للاستخدام في مسألة إعادة التعرف سنقوم فيما يلي تفصيل النموذج ومراحل العمل عليه

يركّز هذا الفصل على عرض التطوّر المرحلي لنموذج KinematicGNN، بدءًا من النموذج الأساسي وصولًا إلى النسخ المحسّنة التي تتضمن استراتيجيات متقدمة في التعلم المتري، وفصل السمات، والتفكيك العدائي للمعلومات الوضعية. ولا يقتصر العرض على وصف المعمارية النهائية، بل يتناول أيضًا المسار التجريبي الذي قاد إلى هذه الخيارات التصميمية، مدعومًا بدراسات تفكيك (Ablation Studies) وتحليل نقدي لنتائج الأداء، بما في ذلك الحالات التي لم تؤدِّ إلى تحسّن ملحوظ أو أدّت إلى تراجع الأداء.

وبذلك، يسعى هذا الفصل إلى تحقيق هدفين متكاملين: أولًا، تقديم توصيف منهجي واضح للإطار المقترح ومبرراته النظرية؛ وثانيًا، توضيح كيف أسهم التطوير التدريجي المدفوع بالنتائج التجريبية في فهم حدود وإمكانات نماذج إعادة التعرّف القائمة على المشية في سياق أنظمة المراقبة الواقعية. ويمهّد هذا العرض، في نهاية المطاف، للانتقال إلى مناقشة أوسع حول دمج تمثيلات المشية مع مصادر سمات أخرى، وتقييم جدوى هذه المقاربات ضمن بيئات حوسبة محدودة الموارد.

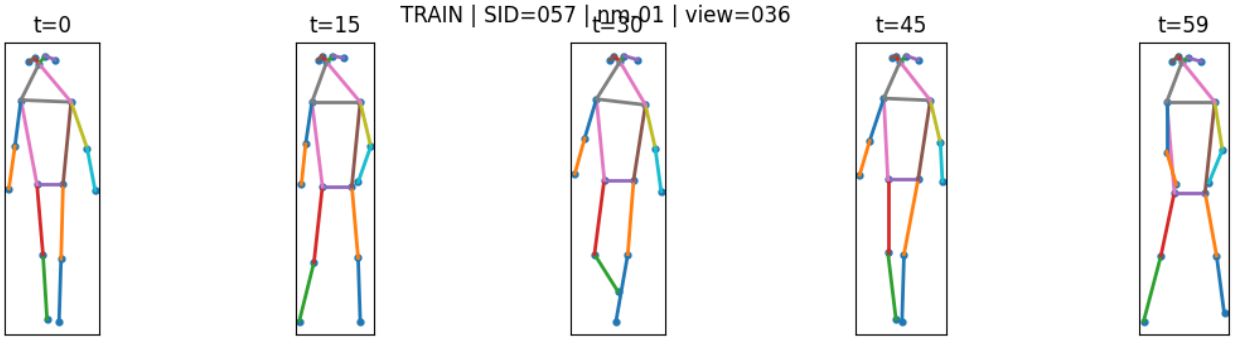
4.4.1. تمثيل البيانات ومدخلات النموذج

يعتمد نموذج KinematicGNN في هذا العمل على تمثيل بنيوي للحركة البشرية مستخلص من بيانات الوضعية (Pose Estimation)، حيث يُمثَّل جسم الإنسان باستخدام نموذج COCO-17 القياسي، الذي يصف الجسم عبر سبعة عشر مفصلًا رئيسيًا تغطي الأطراف العلوية والسفلية والجذع والرأس. وقد تم اختيار هذا التمثيل لكونه واسع الانتشار في نماذج استخراج الوضعية الحديثة، ولتوازنه بين الغنى البنيوي والكلفة الحسابية.

يمثَّل كل مفصل بثلاث خصائص أساسية، هي الإحداثيان المكانيان (x,y)(x, y)(x,y) إضافة إلى قيمة الثقة (confidence) الناتجة عن نموذج استخراج الوضعية. وعليه، يُمثَّل كل إطار زمني بمتجه بعده الكلي 51 بُعدًا (3×17)(17 \times 3)(3×17). يوفّر هذا التمثيل توصيفًا مضغوطًا للحركة، مع تقليل التأثر بعوامل الإضاءة والمظهر الخارجي التي تُعدّ مصدر إرباك رئيسي في أنظمة إعادة التعرّف التقليدية.

	image_name	nose_x	nose_y	nose_conf	left_eye_x	left_eye_y	left_eye_conf	right_eye_x	right_eye_y	right_eye_conf	...
0	./001-bg-01-000/000001.jpg	153.81847	58.585762	0.956070	154.84587	56.530956	0.968925	152.27736	57.044660	0.933316	...
1	./001-bg-01-000/000002.jpg	153.82982	58.295708	0.938918	154.86467	56.226030	0.958389	152.27757	56.743446	0.932614	...
2	./001-bg-01-000/000003.jpg	153.18575	57.753693	0.928092	154.74284	56.196594	0.936697	152.14767	56.196594	0.951392	...

الشكل 27 عينة من البيانات تشرح الصيغة الجدولية لنسخة CASIA-B بظهر فيها كل نقطة مع إحداثياتها



الشكل 28 التمثيلات البيانية المستخرجة من العقد واحداثياتها فوق صيغة COCO-17 تنتمي الصور إلى عنصر المرور ذاته

يتوافق هذا الاختيار مع ما تمّت مناقشته سابقًا في هذه الأطروحة حول محدودية التمثيلات الصورية في توصيف المشية، حتى في الحالات التي تتضمن لقطات متتالية للشخص نفسه، إذ تظل هذه اللقطات مصنّفة ضمن البيانات المعتمدة على الصورة، ولا تحمل الدلالة الزمنية الحركية اللازمة لنمذجة المشية بوصفها سلوكًا ديناميكيًا.

4.4.2. البنية الزمنية وبناء المتتبعات (Skeleton Tracklets)

لا يقتصر تمثيل البيانات في هذا العمل على الإطارات الفردية، بل يتم تجميعها ضمن متتبعات زمنية (Skeleton Tracklets) بطول ثابت **T=60T = 60T=60** إطارًا، تمثل مقطعًا زمنيًا كاملًا لدورة مشي شبه مكتملة. وقد تم اختيار هذا الطول بعناية ليحقق توازنًا بين احتواء الأنماط الحركية الدورية للمشية، والحفاظ على كفاءة المعالجة والتدريب.

يُعد هذا القرار امتدادًا مباشرًا للنقاش المنهجي الذي تم تقديمه في الفصول السابقة حول الفرق الجوهري بين:

- مجموعات البيانات المعتمدة على الصورة (Image-based Re-ID)،
- ومجموعات البيانات الفيديوية أو الحركية المصمّمة خصيصًا لالتقاط المشية.

فعلى الرغم من أن بعض مجموعات بيانات إعادة التعرّف الصورية تتضمن معرفات لتسلسلات قصيرة مرتبطة بمرور الشخص أمام الكاميرا، إلا أن هذه التسلسلات لا تحمل انتظامًا زمنيًا أو اكتمالًا حركيًا يسمح باستخلاص خصائص المشية بشكل موثوق. في المقابل، يفرض استخدام المتتبعات الزمنية في هذا العمل بنية بيانات تُجبر النموذج على التعلّم من تطور الحركة عبر الزمن بدلًا من الاعتماد على لقطات مجزأة أو لحظية.

4.4.3. تهيئة المدخلات لنمذجة بنيوية–زمنية

بعد استخراج المتتبعات الزمنية، يتم تنظيم البيانات في مصفوفة ذات الأبعاد:

T×N×CT \times N \times CT×N×C

حيث:

- TTT هو عدد الإطارات الزمنية.
- N=17N = 17N=17 عدد المفاصل.
- C=3C = 3C=3 عدد القنوات لكل مفصل.

يتيح هذا التنظيم تفسير كل إطار على أنه رسم بياني (Graph) مستقل، تمثل عقده المفاصل، بينما تُعرَّف الحواف وفق البنية التشريحية للجسم البشري. وبذلك، يصبح من الممكن تطبيق الشبكات العصبية الرسومية لمعالجة العلاقات المكانية داخل الإطار الواحد، تمهيدًا لدمج هذه المخرجات لاحقًا عبر نماذج زمنية قادرة على التقاط ديناميكيات المشي.

ويُعد هذا التمثيل المدخلي حجر الأساس الذي بُنيت عليه المعمارية المقترحة، إذ يسمح بفصل واضح بين النمذجة المكانية والنمذجة الزمنية، ويمهّد لإدخال آليات أكثر تقدّمًا للتحكم في نوعية المعلومات المتعلَّمة، كما سيُناقش في الأقسام اللاحقة من هذا الفصل.

4.4.4. المعمارية الأساسية لنموذج KinematicGNN

النمذجة المكانية باستخدام GCN والنمذجة الزمنية باستخدام Transformer

الدافع العام للتصميم المعماري

كما تم توضيحه في القسم السابق، فإن تمثيل المشية البشرية عبر بيانات الوضعية يفرض طبيعة بيانات ذات بعدين متكاملين: بعد مكاني يصف العلاقات البنيوية بين المفاصل داخل الإطار الواحد، وبعد زمني يعبّر عن تطور هذه العلاقات عبر تسلسل الحركة. ومن هذا المنطلق، لا يُعدّ الاعتماد على نماذج تسلسلية أو مكانية فقط كافيًا لالتقاط الخصائص التمييزية للمشية، الأمر الذي استدعى اعتماد معمارية هجينة تجمع بين نمذجة الرسوم البيانية والنمذجة الزمنية.

يعتمد نموذج **KinematicGNN** على فصل واضح بين هذين البعدين، حيث تُعالج العلاقات المكانية أولًا باستخدام الشبكات العصبية الرسومية (Graph Neural Networks)، ثم تُدمج المخرجات عبر نموذج قائم على آليات الانتباه الذاتي لنمذجة الديناميكيات الزمنية للحركة.

النمذجة المكانية: الشبكات العصبية الرسومية (GCN)

يُفسَّر كل إطار زمني من المتتبعة الحركية بوصفه رسمًا بيانيًا G=(V,E) G = (V, E) G=(V,E)، حيث تمثل العقد VVV المفاصل البشرية، بينما تمثل الحواف EEE الروابط التشريحية الطبيعية بينها. يتيح هذا التمثيل التقاط الاعتماديات المكانية بين المفاصل بطريقة صريحة، بخلاف التمثيلات المتجهية المسطّحة التي تهمل البنية الهيكلية للجسم.

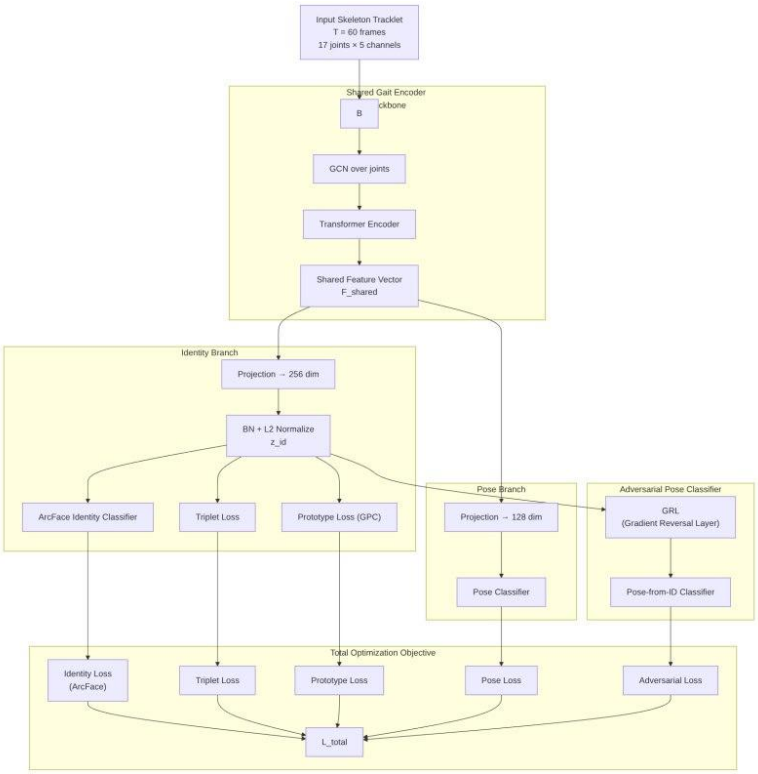
في هذا العمل، تُستخدم عدة طبقات من Graph Convolutional Networks (GCN) لمعالجة هذه الرسوم البيانية، حيث تقوم كل طبقة بتجميع معلومات كل مفصل من جيرانه المباشرين، ما يسمح بانتشار المعلومات البنيوية عبر الهيكل العظمي تدريجيًا. وقد تم اختيار عدد محدود من الطبقات مع أبعاد خفية معتدلة، تحقيقًا لتوازن بين القدرة التمثيلية والاستقرار التدريبي، خصوصًا في ظل محدودية حجم بيانات المشية مقارنةً بمجموعات البيانات الصورية الضخمة.

يُتبع كل تلافيف رسومية بعمليات تطبيع وتفعيل غير خطي، ما يساعد على تقليل التباين بين المتتبعات المختلفة وتحسين قابلية التعميم. ونتيجةً لذلك، ينتج عن هذه المرحلة تمثيل مكاني لكل إطار يعبّر عن الوضعية البنيوية للجسم في تلك اللحظة الزمنية، مع تضمين العلاقات بين المفاصل بشكل ضمني في الفضاء المتعلم.

من الإطارات إلى التسلسل: الحاجة إلى النمذجة الزمنية

على الرغم من أن التمثيلات المكانية المستخرجة بواسطة GCN قادرة على توصيف الوضعية اللحظية للجسم، إلا أنها لا تكفي بمفردها لتمييز المشية، التي تُعد بطبيعتها نمطًا ديناميكيًا دوريًا. فالأفراد قد يتشاركون أوضاعًا متشابهة في لحظات معينة، بينما يكمن الاختلاف الحقيقي في كيفية الانتقال بين هذه الأوضاع عبر الزمن.

لذلك، يتم تجميع التمثيلات المكانية الناتجة عن جميع الإطارات ضمن تسلسل زمني يُمرَّر إلى مرحلة النمذجة الزمنية، حيث يصبح الهدف هو التقاط العلاقات طويلة المدى، والتغيرات الدورية، والاعتماديات الزمنية غير المحلية التي تميّز نمط مشي شخص عن آخر.



الشكل 29 معمارية KINEMATICGNN

النمذجة الزمنية باستخدام Transformer Encoder

لاستخلاص الخصائص الزمنية للمشية، يعتمد نموذج KinematicGNN على Transformer Encoder، الذي أثبت فعاليته في نمذجة التسلسلات طويلة الأمد دون القيود المتأصلة في النماذج التكرارية التقليدية. تقوم آلية الانتباه الذاتي داخل الـ Transformer بحساب أهمية كل إطار بالنسبة إلى بقية الإطارات في المتتبعة، ما يسمح للنموذج بالتركيز على المقاطع الأكثر تمييزًا في دورة المشي.

يمكّن هذا التصميم النموذج من:

- التقاط الأنماط الدورية للمشية،
- التعامل مع اختلاف سرعة الحركة بين الأفراد،
- وتخفيف أثر الضجيج أو الإطارات غير المستقرة الناتجة عن أخطاء استخراج الوضعية.

وبذلك، يتحول تسلسل التمثيلات المكانية إلى تمثيل زمني غني يعكس السلوك الحركي الكلي للشخص بدلًا من الاعتماد على لحظات منفصلة.

الشعاع المشترك للخصائص (Shared Feature Representation)

ينتج عن مرحلة Transformer متجه خصائص مشترك ويُعد هذا المتجه حجر الأساس في بقية المعمارية. إذ يحتوي هذا التمثيل على مزيج من:

- معلومات بنيوية مرتبطة بشكل الجسم،
- معلومات زمنية مرتبطة بديناميكيات المشي،
- ومؤشرات ضمنية لكل من الهوية والوضعية.

غير أن هذا التداخل في المعلومات، على الرغم من ضرورته في المراحل الأولى من التعلّم، يطرح تحديًا جوهريًا يتمثل في صعوبة فصل السمات التمييزية للهوية عن السمات المرتبطة بالوضعية أو زاوية الالتقاط. ومن هنا، تبرز الحاجة إلى تصميم معماري لاحق يسمح بالتحكم الصريح في نوعية المعلومات التي يتم الاحتفاظ بها أو قمعها في التمثيل النهائي، وهو ما سيتم تناوله بالتفصيل في القسم التالي المتعلق بالتصميم متعدد الفروع وفصل السمات.

محدودية التمثيل المشترك وتداخل خصائص المشية والوضعيّة

بعد أن صممنا المعمارية الأساسية (GCN + Transformer) وأنتجت شعاع مشتركًا للخصائص لاحظنا أن هذا التمثيل رغم غناه لا يضمن أن النموذج يتعلّم ''هوية الشخص'' فعليًا بالمعنى الذي نريده في إعادة التعرّف القائمة على المشية. السبب أن الشعاع المشترك يجمع في آنٍ واحد معلومات عديدة: بنية الجسم، ديناميكيات الحركة عبر الزمن، لكن أيضًا إشارات مرتبطة بالوضعية وزاوية الالتقاط وأحيانًا ضجيج ناتج عن أخطاء استخراج الوضعية. بمعنى آخر، وجدنا أن النموذج قد يحقق نتائج جيدة أثناء التدريب لأنه يستفيد من قرائن ''سهلة'' مثل الوضعية، لكنه عند تغيير الزوايا أو الحالات يصبح أداؤه أقل ثباتًا، لأن ما تعلّمه ليس هوية خالصة بل خليط من عوامل متداخلة.

انطلاقًا من هذا الفهم، تكوّن لدينا حدس أساسي: إذا لم نُجبر النموذج على تنظيم معلوماته، فسيمزج الهوية مع الوضعية داخل نفس التمثيل. وبما أن هدفنا هو استخراج تمثيل يصلح لإعادة التعرّف عبر اختلاف الوضعيات، فقد أصبح من الضروري أن ندفع النموذج إلى تمييز ما هو ''هوية'' عمّا هو ''وضعية/زاوية'' بدل ترك الأمر للتعلّم العفوي. ومن هنا جاءت خطوة الانتقال من ''تمثيل مشترك واحد'' إلى ''تصميم متعدد الفروع'' يوزع الأدوار ويقلّل التداخل.

أول ما قمنا به هو إنشاء فرع هوية (Identity Branch) يكون هو المسار الرسمي الذي نريد منه التمييز بين الأشخاص. أخذنا الشعاع المشترك وأسقطناه إلى فضاء أصغر (256 بعدًا)، ثم قمنا بتطبيق تطبيع (BN ثم L2) للحصول على تمثيل الفكرة هنا بسيطة: نحن نريد تمثيلًا ''نظيفًا'' ومستقرًا يسهل قياس التشابه فيه، ويكون مناسبًا للتعلم المتري وخسائر الفصل بين الهويات. عمليًا، هذا الفرع هو الذي سيحمل مسؤولية أداء Re-ID.

لكننا أدركنا أيضًا أن مجرد تخصيص فرع للهوية لا يمنع أن تتسلّل معلومات الوضعية إليه. ولذلك جاءت الخطوة الثانية: بدل أن نحاول ''إلغاء الوضعية'' أو تجاهلها، قررنا أن نعطيها مسارًا صريحًا. أنشأنا فرع وضعية (Pose Branch) يستقبل نفس ولكن يُسقطه إلى بعد أصغر (128 بعدًا) ثم يمرره إلى مصنّف وضعية. حدسنا هنا كان أن: الوضعية ستظهر في التعلم على أي حال؛ فإذا لم نخصص لها مسارًا واضحًا فقد تتكدّس داخل تمثيل الهوية وتفسده. أما إن جعلناها ''معلومة متعلمة في مكانها''، فنحن نسحب جزءًا من العبء بعيدًا عن فرع الهوية ونزيد قابلية السيطرة على ما يتعلمه كل جزء.

رغم ذلك، بقيت مشكلة ثالثة: حتى مع وجود فرع وضعية، ما زال من الممكن أن يظل يحمل معلومات وضعية كافية، لأن النموذج قد يجد في ذلك منفعة لتحسين الفصل أثناء التدريب. لذلك كانت الخطوة التي مثّلت نضج الفكرة: التفكيك العدائي (Adversarial Disentanglement). أضفنا مصنّفًا إضافيًا يحاول التنبؤ بالوضعية انطلاقًا من ولكننا وضعنا بينهما طبقة GRL التي تعكس التدرّج أثناء التدريب. بهذه الطريقة يحدث شيء مقصود: المصنّف يحاول كشف الوضعية من تمثيل الهوية، بينما النموذج (بسبب GRL) يُدفَع لِ جعل ذلك صعبًا عبر إزالة الإشارات الوضعية . بعبارة أبسط: نحوّل وجود الوضعية داخل تمثيل الهوية إلى ''شيء يُعاقَب عليه'' بدل أن يكون شيئًا يُكافأ ضمنيًا.

بهذا التسلسل، تطوّر التفكير من ملاحظة محدودية التمثيل المشترك، إلى بناء مسارات متخصصة، ثم إلى فرض فصل أقوى عبر آلية عدائية. ويمكن تلخيص منطق التطوير كالتالي:

- وجدنا أن تمثيلًا واحدًا غنيًا لا يعني بالضرورة تمثيلًا صحيحًا للهوية.
- افترضنا أن مصدر المشكلة هو اختلاط الهوية مع الوضعية داخل نفس الفضاء.
- لذلك صممنا فرع هوية واضحًا، وخصصنا للوضعية فرعًا مستقلًا بدل تركها تتسرّب.

- ثم عززنا الفصل بإضافة GRL كي تصبح ''الوضعية داخل تمثيل الهوية'' أمرًا غير مرغوب تدريبيًا.

هذا المسار لا يقدّم معمارية نهائية فقط، بل يوضح كيف أن كل تعديل كان استجابة مباشرة لحدسٍ مدعوم بالملاحظة: نريد تمثيل هوية ثابتًا عبر تغيّر الوضعية، وبالتالي كان لا بد من هندسة معمارية تُجبر النموذج على هذا الفصل بدل أن نأمل حدوثه تلقائيًا.

هندسة توابع الخسارة ودورها في دعم الفصل المعماري

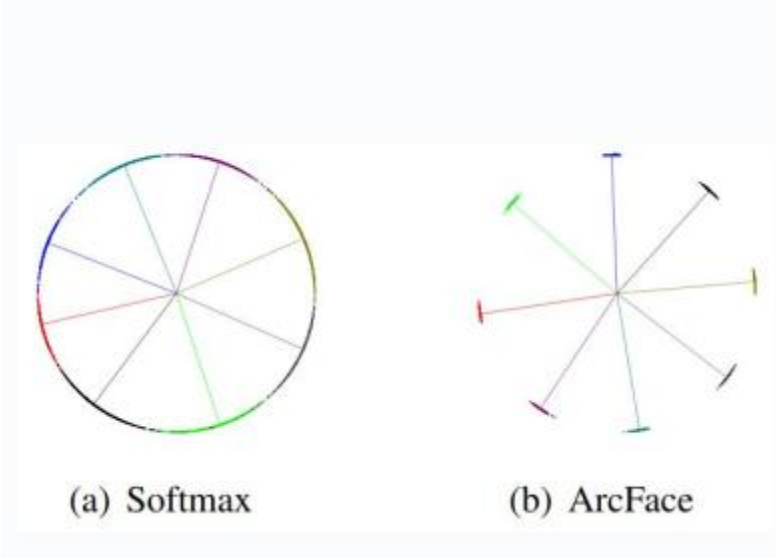
بعد الانتقال من التمثيل المشترك إلى التصميم متعدد الفروع، لم يعد كافيًا الاعتماد على دالة خسارة واحدة أو عامة لتوجيه عملية التعلّم. إذ إن وجود فروع متخصصة للهوية والوضعية، إضافة إلى آلية التفكيك العدائي، يفرض بالضرورة أن تكون توابع الخسارة متوافقة مع الأدوار الوظيفية لكل فرع. ومن هنا، لم تُستخدم الخسائر في هذا العمل بوصفها أدوات تحسين رقمية فقط، بل كآليات توجيه صريحة تضمن أن يتعلّم كل فرع النوع الصحيح من الخصائص.

خسائر فرع الهوية: تعزيز التمييز دون الاعتماد على قرائن غير مستقرة

في فرع الهوية، كان الهدف الأساسي هو دفع النموذج إلى تعلّم تمثيل يميّز بين الأشخاص اعتمادًا على خصائص المشية، لا على إشارات وضعية أو هندسية مؤقتة. ولتحقيق ذلك، تم الاعتماد على مزيج من ثلاث خسائر متكاملة، لكل منها دور مختلف.

أولًا، استُخدمت Triplet Loss لتشكيل الفضاء التمثيلي على مستوى المسافات، بحيث تُقَرَّب العينات التابعة للشخص نفسه وتُبعَد عن عينات الأشخاص الآخرين. هذه الخسارة وفّرت أساسًا للتعلّم المتري، لكنها بمفردها لا تفرض بنية قوية على الفضاء، وقد تسمح للنموذج بالاعتماد على أي إشارات تساعده على الفصل، بما في ذلك إشارات غير مرغوبة.

لذلك، تم إدخال ArcFace Loss بوصفها خسارة زاوية تفرض فصلًا أوضح بين الهويات على مستوى الاتجاهات في الفضاء المعياري. ساعدت هذه الخسارة على جعل حدود الفصل أكثر صرامة، لكنها لم تكن كافية وحدها لمعالجة مشكلة التداخل بين الهوية والوضعية، كما أظهرت بعض التجارب التي لم تحقق تحسنًا جوهريًا عند الاعتماد عليها فقط.



الشكل 30 يوضح الشكل كيف بفصل ARCFACE الفضاء الهوياتي بشكل أكثر وضوحاً من SOFTMAX

أما Prototype Loss (GPC)، فقد أُضيفت استجابةً لملاحظة أن عينات الشخص الواحد قد تتوزع في الفضاء التمثيلي بسبب اختلاف الأوضاع ودورات المشي. هدفت هذه الخسارة إلى سحب تمثيلات الشخص الواحد نحو ممثل (prototype) مشترك، ما عزّز تماسك التمثيل الهوياتي عبر المتتبعات المختلفة، وساهم في تحسين الاستقرار، خاصة على مستوى mAP.

بهذا المزيج، لم يعد فرع الهوية يكتفي بالفصل بين الأشخاص، بل أصبح موجَّهًا لبناء تمثيل متماسك وموزّع بشكل منظم، وهو ما يتماشى مع الهدف المعماري للفصل بين الهوية وبقية العوامل.

خسارة فرع الوضعية: سحب المعلومات غير المرغوبة إلى مسارها الخاص

في المقابل، لم يكن الهدف من فرع الوضعية تحسين أداء إعادة التعرّف بشكل مباشر، بل توفير مخرج صريح للمعلومات الوضعية التي لو تُركت دون توجيه لتسرّبت إلى تمثيل الهوية. ولهذا الغرض، استُخدمت Pose Classification Loss لتشجيع هذا الفرع على تعلّم تمثيل قادر على توصيف الوضعية أو حالة المشي بشكل واضح.

أدّى وجود هذه الخسارة إلى إعادة توزيع الأدوار داخل النموذج: فبدل أن تضطر شبكة الهوية إلى الاحتفاظ بمعلومات وضعية "للاستفادة منها"، أصبح فرع الوضعية هو المكان الطبيعي لتعلّم هذه المعلومات. وبذلك، دعمت خسارة الوضعية الهدف المعماري القائل بأن الوضعية يجب أن تُنمذج، لكن خارج فضاء الهوية.

الخسارة العدائية: جعل تسرّب الوضعية إلى تمثيل الهوية أمرًا غير مرغوب

رغم الفصل البنيوي بين الفروع، ظل احتمال تسرّب بعض المعلومات الوضعية إلى تمثيل الهوية قائمًا. ولهذا السبب، جاءت الخسارة العدائية المرتبطة بطبقة **GRL** لتلعب دورًا حاسمًا في دعم الفصل المعماري.

في هذا الإطار، يحاول مصنّف عدائي التنبؤ بالوضعية انطلاقًا من تمثيل الهوية، بينما تقوم GRL بعكس إشارة التدرّج أثناء التدريب. ونتيجةً لذلك، يصبح وجود معلومات وضعية داخل شعاع عاملًا سلبيًا، إذ يؤدي إلى زيادة الخسارة بدل تقليلها. وبذلك، لا يكتفي النموذج بتعلّم تمثيل هوية تمييزي، بل يُدفَع أيضًا إلى جعله غير قابل للتنبؤ الوضعي.

أظهرت التجارب أن هذه الخسارة حسّاسة بطبيعتها، وأن الإفراط في وزنها قد يؤدي إلى كبح التعلّم التمييزي نفسه، وهو ما يفسّر حالات تراجع الأداء عند تشديد الفصل العدائي. وتؤكد هذه النتيجة أن الخسارة العدائية ليست هدفًا بحد ذاتها، بل أداة دقيقة يجب موازنتها بعناية ضمن الإطار الكلي للتعلّم.

الخسارة الكلية: تجسيد للانسجام بين المعمارية والتعلّم

تُجمع جميع هذه الخسائر ضمن دالة خسارة كلية موزونة، بحيث يعكس وزن كل خسارة الدور الوظيفي للفرع الذي تنتمي إليه. وبهذا، تصبح عملية التدريب تجسيدًا مباشرًا للتصميم المعماري:

- خسائر الهوية تدعم بناء فضاء تمييزي مستقر،
- خسارة الوضعية تسحب المعلومات الهندسية إلى مسارها الخاص،
- والخسارة العدائية تفرض قيدًا يمنع تداخل الأدوار بين الفروع.

وبذلك، لا يمكن فصل هندسة دوال الخسارة في هذا العمل عن المعمارية نفسها، إذ إن كل خسارة صُمّمت لتدعم قرارًا معماريًا محددًا، وكل قرار معماري احتاج بدوره إلى خسارة مناسبة ليترسّخ أثناء التعلّم.

معمارية نموذج KinematicGNN

يعتمد نموذج KinematicGNN على تمثيل بنيوي–زمني لإعادة التعرّف القائمة على المشية باستخدام متتبعات هيكلية بطول T=60T=60T=60 إطارًا، حيث يُمثَّل كل إطار عبر 17 مفصلًا بعدة قنوات توصّف موضع المفصل وموثوقيته. يهدف النموذج إلى استخلاص تمثيل هوياتي تمييزي مع التحكم الصريح في تداخل معلومات الوضعية.

في المرحلة الأولى، تُعالج المفاصل داخل كل إطار باستخدام شبكات عصبية رسومية (GCN) لالتقاط العلاقات التشريحية المكانية بين المفاصل. تُمرَّر المخرجات بعد ذلك إلى Transformer Encoder لنمذجة الديناميكيات الزمنية للمشية، ما يسمح بالتقاط الأنماط الدورية والعلاقات طويلة المدى بين الإطارات. ينتج عن هذا الجزء شعاع خصائص مشترك يجمع معلومات حركية وبنيوية ووضعية.

لمعالجة تداخل السمات داخل هذا التمثيل المشترك، يعتمد النموذج تصميمًا متعدد الفروع. في فرع الهوية، يُسقَط الشعاع المشترك إلى فضاء بعده 256، ويُطبَّع باستخدام Batch Normalization وL2 Normalization لإنتاج تمثيل هوياتي، يُدرَّب باستخدام مزيج من خسائر ArcFace وTriplet وPrototype لتعزيز الفصل والتماسك الهوياتي. وبالتوازي، يُمرَّر شعاع إلى فرع الوضعية، حيث يُنمذج بشكل صريح في فضاء بعده 128 باستخدام خسارة تصنيف وضعية، بهدف سحب المعلومات الوضعية خارج فضاء الهوية.

ولضمان عدم تسرّب معلومات الوضعية إلى تمثيل الهوية، يُضاف مكوّن تفكيك عدائي يعتمد على طبقة عكس التدرّج (GRL)، حيث يحاول مصنّف عدائي استنتاج الوضعية بينما يُجبر النموذج الأساسي على جعل هذا التنبؤ صعبًا. تُجمع جميع الخسائر ضمن دالة خسارة كلية واحدة، بما يحقق انسجامًا بين المعمارية واستراتيجية التعلّم.

يمكّن هذا التصميم من الانتقال من تمثيل مشترك غير مضبوط إلى إطار موجَّه يوزّع الأدوار بين المكوّنات المختلفة، ويعزّز استخلاص تمثيل هوياتي قائم على المشية يتمتع بثبات أعلى وقابلية أفضل للتعميم في سيناريوهات المراقبة الواقعية.

النتائج:

سنستعرض فيما يلي تطور النتائج بالترافق مع تطور معمارية KinematicGNN

Exp	Model Variant	Key New Idea	Rank-1 (Test)	Rank-5	mAP (Test)	Val Rank-1	Val mAP

≈ 0.307	≈ 0.960	0.1565	0.9762	0.9216	Triplet + CE + GPC (default λ)	**Baseline**	Exp 1
0.314	0.968	0.1826	**0.9804**	**0.9274**	↑Positives per ID (P=12, K=6), λ(GPC)=0.3, margin=0.35	**Tuned Metric-Learning**	Exp 2
0.307	0.961	0.1819	0.9773	0.9213	Angular-margin Softmax instead of linear CE	**+ArcFace head**	Exp 3
0.314	0.964	**0.1891**	0.9774	0.9187	ID branch + Pose branch + Adversarial Pose Classifier	**+Feature Disentanglement**	Exp 4
0.273	0.939	0.1585	0.958	0.8723	grlλ=2, λ(pose_adv)=0.7, MI orthogonality loss	**Strong Adv + MI penalty**	Exp 5b

الجدول 18 تطور معمارية KIENMATICGNN والنمتائج في كل مرحلة

التحليل الكمي للنتائج (Quantitative Analysis)

اعتمد تقييم الأداء على المقاييس الشائعة في مهام إعادة التعرّف، وبالأخص Rank-1 Accuracy و Mean Average Precision (mAP). وقد أظهرت النتائج اختلافًا واضحًا في حساسية كل مقياس للتعديلات المعمارية المختلفة.

من جهة، حقّق النموذج الأساسي أداءً مقبولًا على مستوى Rank-1، ما يشير إلى قدرته على مطابقة بعض العينات الصحيحة في المرتبة الأولى. غير أن قيمة mAP كانت منخفضة نسبيًا، وهو ما يعكس ضعف تماسك التمثيل الهوياتي، خصوصًا عند وجود اختلافات وضعية أو زمنية داخل هوية الشخص الواحد. ويشير هذا السلوك إلى أن النموذج كان يعتمد جزئيًا على إشارات غير مستقرة لتحقيق المطابقة الأولى، دون بناء فضاء تمثيلي متماسك على مستوى الاسترجاع الكامل.

عند تحسين استراتيجية التعلم المتري، لوحظ تحسّن متزامن في Rank-1 وmAP، ما يؤكد أن تنظيم العينات داخل الفضاء التمثيلي يسهم في تقليل التشتت الداخلي. إلا أن هذا التحسّن ظل محدودًا، الأمر الذي أشار إلى أن المشكلة لا تتعلق فقط بكيفية توزيع العينات، بل بنوعية الخصائص التي يتم تمثيلها أصلًا.

أما إدخال ArcFace، فقد حسّن حدود الفصل بين الهويات على المستوى الزاوي، لكنه لم يُترجم إلى قفزة نوعية في mAP. وتدل هذه النتيجة على أن الفصل الأقوى بين الهويات لا يكفي إذا كان الفضاء نفسه يحتوي على خصائص غير ثابتة، مثل الوضعية أو زاوية الالتقاط.

التحسن الأوضح في mAP ظهر عند إدخال فرع الوضعية، حيث ارتفعت قيمة هذا المقياس بشكل ملحوظ مقارنةً بالمراحل السابقة. ويُعد هذا الارتفاع مؤشرًا قويًا على أن الفصل البنيوي للسمات ساهم في زيادة تماسك تمثيلات الهوية عبر المتتبعات المختلفة، حتى عندما لم تتحسن قيمة Rank-1 بنفس النسبة. ويؤكد ذلك أن النموذج أصبح أكثر قدرة على استرجاع جميع العينات الصحيحة، لا مجرد المطابقة الأولى.

أخيرًا، أظهرت التجارب المرتبطة بالتفكيك العدائي أن التأثير الكمي لهذا المكوّن يعتمد بشدة على درجة تفعيله. ففي الإعدادات المعتدلة، لوحظ تحسّن إضافي في mAP، ما يدل على تقليل تسرّب المعلومات الوضعية إلى تمثيل الهوية. أما عند تشديد الفصل العدائي، فقد انخفض الأداء، وهو ما يعكس فقدان النموذج لبعض الخصائص البنيوية المفيدة للتمييز.

التحليل النوعي لسلوك النموذج (Qualitative Analysis)

يوفّر التحليل النوعي تفسيرًا أعمق لما تعكسه القيم الرقمية من سلوك تعلّمي. فقد أظهرت النماذج المبكرة ميلًا إلى التمييز اعتمادًا على أوضاع مشي أو زوايا رؤية متشابهة، حيث كانت الأخطاء تتركّز غالبًا في الحالات التي يتغيّر فيها اتجاه المشي أو تتبدّل زاوية الكاميرا بشكل واضح. ويشير ذلك إلى أن النموذج كان يخلط بين الهوية والوضعية داخل نفس الفضاء التمثيلي.

بعد إدخال فرع الوضعية، أصبح هذا النوع من الأخطاء أقل تكرارًا، إذ بات النموذج أكثر قدرة على مطابقة المتتبعات التي تمثل نفس الشخص في أوضاع مشي مختلفة. ويُفسَّر هذا السلوك بأن المعلومات الوضعية لم تعد تؤثّر بشكل مباشر على قرار المطابقة الهوياتي، بل أصبحت ممثَّلة في مسار مستقل.

أما التفكيك العدائي، فقد أظهر أثرًا نوعيًا يتمثل في تقليل اعتماد النموذج على قرائن هندسية واضحة، مثل اتجاه الجسد أو تماثل الوضعية بين المتتبعات. غير أن الإفراط في هذا الفصل أدى إلى حالات فشل يصبح فيها النموذج غير قادر على التمييز بين أشخاص ذوي بنية جسدية متقاربة وأنماط مشي متشابهة، ما يوضح أن بعض الخصائص البنيوية—وإن لم تكن خالصة للمشية—تظل ضرورية للتمييز.

تقييم نهائي للنموذج

تحسينات التعلم المتري تؤثر بشكل مباشر على دقة المطابقة الأولى، بينما يسهم الفصل البنيوي للسمات في تحسين تماسك التمثيل الهوياتي عبر المتتبعات المختلفة، وهو ما انعكس بوضوح في تحسّن قيمة mAP. كما أظهرت النتائج أن التفكيك العدائي

يمكن أن يعزز هذا الفصل عند استخدامه باعتدال، في حين يؤدي الإفراط فيه إلى كبح التعلّم التمييزي نفسه، ما يبرز وجود نقطة توازن دقيقة بين الفصل والحفاظ على الخصائص البنيوية المفيدة.

وتُفسَّر القيم العددية للمقاييس، ولا سيما mAP، ضمن سياق الإعداد التجريبي المعتمد في هذا العمل، والذي يتعامل مع تمثيلات هيكلية مستخرجة آليًا من بيانات مراقبة واقعية غير مثالية. ومن هذا المنطلق، لا يُعد انخفاض القيم المطلقة مؤشرًا على ضعف النموذج، بقدر ما يعكس تعقيد المشكلة وصرامة الشروط التي تم فيها التقييم.

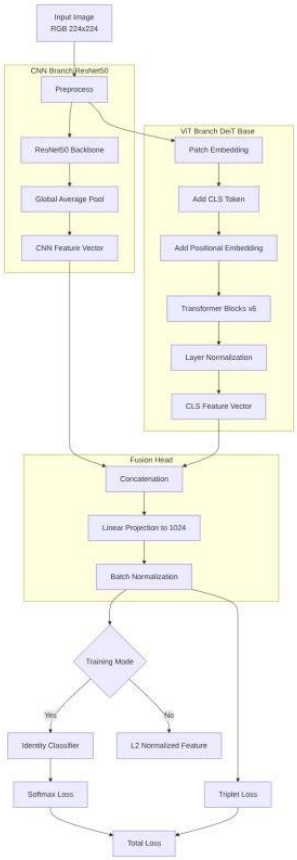
4.5. النموذج الهجين المعتمد على المظهر Hybrid Appearance Transformer (HAT-ReID)

ضمن هذا القسم في في الإطار العملي قمنا بالاعتماد بشكل اساسي على إطار العمل torchreid الذي قمنا بتفصيله سابقاً في الدراسة المرجعية

إن كون المسألة هنا تشبه في جوهرها مسائل استرجاع صورية كلاسيكية بمعنى تشفير صور البيانات واستخراج الشعاع الممثل لصورة جديدة ومقارنتها بأشعة لصورة التي يعلمها النظام سابقاً فقد قمنا بتجربة الحلول الكلاسيكية ضمن عدة تجارب تشمل ResNet50 وكذلك OSNet وذلك مع مجموعة بيانات Market1501.

بعد دراسة النموذج الهجين الذي قمنا بالتطرق إليه في الدراسة المرجعية والذي اعتمد بنية هجينة من mobilenet لاستخراج الخصائص ومحول بصري مقتطع (أربع طبقات) TE-TransReID.

بعد محاولة تكرار النتائج التي طرحها هذا النموذج قمنا بالاستئناس به في محاولة لبناء نموذج هجين له الروح ذاتها ولكن مع اعتماد ResNet بالإضافة إلى محول بصري مقتطع ولكن مؤلف من 6 طبقات.



الشكل 31 معمارية النموذج الهجين المعتمد على المظهر

وقمنا في النهاية بمقارنة جميع النتائج المتعلقة بهذه التجارب والموضحة جميعاً في الجدول التالي.

Rank-20	Rank-10	Rank-5	Rank-1	mAP	Model
%98.00	%96.80	%94.80	%88.40	%70.30	OSNet
%98.00	%96.70	%95.00	%88.20	%73.20	ResNet50 (60 epochs)
%97.50	%96.30	%94.40	%86.00	%68.70	Swin (120 epochs)
%98.40	%97.70	%96.60	%90.70	%76.20	Hybrid Model (ResNet50 + 6-layer ViT)
%97.20	%95.20	%92.10	%80.30	%56.40	MobileNet + 4-layer ViT

الجدول19 مقارنة نتائج تحارب النماذج المعتمكدة على المظهر ومن ضمنها نموذجنا HAT

تُظهر النتائج بوضوح أن النموذج الهجين المقترح يحقق أفضل أداء بين جميع النماذج التي تم اختبارها، سواء على مستوى mAP أو Rank-1، مما يؤكد جدوى الجمع بين التمثيل المحلي القوي الذي توفره الشبكات الالتفافية العميقة، والقدرة على نمذجة العلاقات العالمية التي يتيحها المحوّل البصري المقتطع. وفي المقابل، تُبرز النتائج المحدودة لنموذج MobileNet + ViT أن خفة النموذج، رغم فائدتها من حيث الكلفة الحسابية، تأتي على حساب القدرة التمييزية في سياق إعادة التعرّف.

ومع ذلك، تؤكد هذه النتائج أيضًا ما تم التوصل إليه في الأدبيات الحديثة، وهو أن مسألة إعادة التعرّف المعتمدة حصريًا على المظهر ومن صورة واحدة قد بلغت درجة عالية من النضج البحثي، حيث أصبحت التحسينات المتحققة غالبًا تدريجية ومحدودة. وعليه، فإن أهمية هذه التجارب لا تكمن في تحقيق طفرة رقمية بحد ذاتها، بل في إثبات القدرة على مجاراة ما توصلت إليه الأبحاث الحديثة، وفهم عميق لآلياتها وحدودها، بما يمهّد بصورة منطقية للانتقال في الفصول اللاحقة إلى مقاربات أكثر ثراءً، تعتمد دمج مصادر معلومات إضافية مثل المشية والوضعية الحركية.

4.6. النموذج الوصفي الدلالي

4.6.1. التوصيف النصي لصور المشاة كخاصية دلالية

قمنا بعدة تجارب تتعلق بالاستفادة من النماذج اللغوية البصرية بهدف استثمارها في تطبيقات إعادة التعرف.

فيما يلي مثلا محاولة استقراء مدى قدرة نموذج Qwen2.5VL: 4B على التوصيف لصور المارة، في يلي عدد من الصور المنتقاة من مجموعة البيانات ICFG-PEDES ومقارنة التوصيف المذكور فيها بما نتج عن تلقين أمر توصيف للنموذج المذكور للصورة ذاتها، وفقاً لأمر التلقين التالي:

Describe this pedestrian image in detail (assuming the pedestrian is walking).

قمنا بتجربة دراسة قابلية نماذج Vision-Language Models (VLMs) لتوليد أوصاف نصية ذات قيمة تعريفية عالية تخدم مسألة Person Re-Identification (ReID)، وهي مسألة تتطلب تمثيلًا دلاليًا دقيقًا للسمات البصرية المستقرة التي تسمح بتمييز هوية الشخص عبر مشاهد وكاميرات مختلفة. وعلى خلاف مهام Generic Image Captioning التي تركز على الوصف العام أو السردي للمحتوى البصري، يتطلب سياق ReID أوصافًا موجهة ومنظمة تُبرز عناصر مثل نوع الملابس وألوانها، الإكسسوارات، الوضعية والحركة، والخصائص الجسدية المميزة.

عينات من مجموعة البيانات ICFG-PEDES		
A man with short black hair is wearing a red insulated jacket with a hood and a pair of blue denim jeans. He is wearing yellow running shoes with white soles and is carrying a black backpack and has both his hands in the pocket.	A woman in her twenties with shoulder-length black hair is wearing a black shirt layered with a hooded grey insulated jacket. She is also wearing a black skirt and is carrying a black backpack	A young woman with long black hair is wearing a blue puffer parka jacket with black leggings and black knee-high boots. She is also carrying a blue and white backpack with red straps.

قمنا بتصميم تجربة كالتالي:

اختبار نموذجين لغويين مختلفين والقيام بمحاولة كشف مدى التشابه بين وصفهما لصور المارة وذلك من أجل حالتين لأوامر التلقين: بسيط ومعقد

الأمر المعقد كان كالتالي:

Focus strictly on identity-relevant visual attributes.

Your description must include:

- Gender and approximate age group
- Upper-body clothing (type, color, pattern, and material if visible)
- Lower-body clothing (type and color)
- Footwear (type and color)
- Accessories (e.g., backpack, handbag, hat, glasses, carried objects)
- Pose and action (e.g., walking, standing, direction of movement)
- Any distinctive physical or visual characteristics

Avoid mentioning irrelevant background details.

Do not speculate beyond what is visually observable.

Use clear, explicit, and complete sentences.

وهو موجه خصيصًا لمهام Person Re-Identification، بهدف دفع نموذج Vision-Language Model إلى توليد أوصاف نصية غنية ومنضبطة تركز على السمات البصرية ذات القيمة التعريفية. وعلى خلاف الـ Simple Prompt الذي يكتفي بطلب وصف عام للصورة، يفرض الـ Complex Prompt مجموعة صريحة من القيود والتعليمات التي تُجبر النموذج على تغطية محاور دلالية محددة تشمل نوع الملابس وألوانها، الإكسسوارات، الوضعية، والحركة، إضافةً إلى السمات الجسدية المميزة.

كانت نتائج التجربة كالتالي:

Overall Performance Ranking

1. LLaVA-1.5-7B + complex prompt
   - Composite Score: 0.4110
   - ROUGE-L: 0.1890
   - Semantic Similarity: 0.5032

- BLEU-4: 0.0168
- Error Rate: 0.0%

2. LLaVA-1.5-7B + simple prompt
   - Composite Score: 0.4060
   - ROUGE-L: 0.2405
   - Semantic Similarity: 0.4281
   - BLEU-4: 0.0271
   - Error Rate: 0.0%
3. Qwen2.5-VL-3B-Instruct + simple prompt
   - Composite Score: 0.4052
   - ROUGE-L: 0.2233
   - Semantic Similarity: 0.4400
   - BLEU-4: 0.0311
   - Error Rate: 0.0%
4. Qwen2.5-VL-3B-Instruct + complex prompt
   - Composite Score: 0.4038
   - ROUGE-L: 0.1742
   - Semantic Similarity: 0.4934
   - BLEU-4: 0.0176
   - Error Rate: 0.0%

الجدول 20 مقارنة النموذجين البصريين QWEN و LLAVAمع نمطي التلقين البسيط والمعقد

أظهرت نتائج التقييم أن أداء نماذج Vision-Language Models يتأثر بشكل مباشر بالتفاعل بين معمارية النموذج وتصميم الـ prompt. ووفقًا لمقياس Composite Score، حقق نموذج LLaVA-1.5-7B أفضل أداء إجمالي عند استخدام كلٍ من complex وsimple prompt، متقدمًا بفارق طفيف على نموذج Qwen2.5-VL-3B-Instruct، مع تقارب واضح في القيم العددية بين جميع الإعدادات، ما يشير إلى عدم وجود تفوق مطلق لأي منها.

كما بيّنت النتائج أن استخدام complex prompt يعزز الاتساق الدلالي للأوصاف المولّدة، كما يظهر في ارتفاع قيم Semantic Similarity، إلا أنه يؤدي في المقابل إلى انخفاض في مقاييس ROUGE-L وBLEU-4، مما يعكس تحررًا

أكبر من النص المرجعي. وعلى العكس، يحقق simple prompt تطابقًا نصيًا أعلى، خاصةً مع نموذج Qwen2.5-VL، ولكن على حساب الكثافة الدلالية.

تعكس هذه المفاضلة توترًا بنيويًا بين التقييم المعجمي والدلالي، وهو توتر جوهري في سياق Person Re-Identification، حيث تُعد القدرة على تمثيل السمات التمييزية أهم من التطابق الحرفي. كما أظهرت جميع الإعدادات استقرارًا تشغيليًا كاملًا مع معدل خطأ صفري، مما يؤكد موثوقية الإطار التجريبي ويهيئ للانتقال إلى مناقشة أعمق حول دمج السمات اللغوية والبصرية في أنظمة ReID متعددة الوسائط.

المعايرة الدقيقة والنماذج اللغوية الصغيرة

بعد أن تبين لنا أن القدرة التوصيفية للنماذج اللغوية في حالتها الأساسية ليست كافية وأن هندس الأوامر لا تحدث أثراً أو تحسينا كاف، وجدنا أنفسنا أمام خيار معايرة دقيقة finetuning كوسيلة لتحسين النتائج.

قمنا بمحاولة تكييف او معايرة دقيقة مع نموذج Qwen2.5_vlm السابق ولكن بعد تقدير الموارد والوقت اللازم لتدريب نموذج مكون من B3 معلمة وتبين لنا أن المقاربة ليست عملية.

لذلك قمنا بالتوجه نحو نماذج لغوية أصغر

- **SmolVLM2**

يُعد SmolVLM2 نموذجًا من فئة Vision-Language Models خفيفة الوزن، صُمِّم خصيصًا لتحقيق توازن عملي بين **القدرة الدلالية متعددة الوسائط** والكلفة الحسابية المنخفضة. بخلاف نماذج LVLM كبيرة الحجم مثل Qwen2-VL أو LLaVA، يستهدف SmolVLM2 سيناريوهات التشغيل الواقعي التي تفرض قيودًا صارمة على الذاكرة وزمن الاستدلال، مع الحفاظ على مستوى مقبول من الفهم البصري–اللغوي. يعتمد النموذج على مُرمِّز بصري مُضغّط مقترن بنموذج لغوي صغير نسبيًا، ما يسمح له بتوليد أوصاف نصية موجزة ودلالية للصور دون الحاجة إلى موارد حوسبة عالية. وتبرز أهمية SmolVLM2 في سياق أنظمة إعادة التعرف على الأشخاص (ReID) بوصفه مرشحًا عمليًا لتوليد تمثيلات دلالية أو Semantic Tokens تُستخدم كمكوّن مساعد للنماذج التمييزية، خصوصًا في البيئات التي لا تسمح بتشغيل LVLMs ضخمة بشكل مستمر. ومع ذلك، فإن محدودية سعة النموذج مقارنةً بالنماذج الأكبر تنعكس على عمق الاستدلال اللغوي وقدرته على التقاط الفروق الدقيقة جدًا في السمات البصرية، ما يجعل استخدامه أكثر ملاءمة كحل عملي خفيف أو كجزء من بنية هجينة بدلًا من كونه بديلًا كاملًا للنماذج الكبيرة.

وكان تركيز التجارب منصبا على smolvlm2 500M على اعتبار أ، حجمه مناسب عملياً للتدريب والمعايرة والتشغيل على الحافة.

النموذج الأساسي (Baseline – دون Fine-tuning)

هدفت هذه التجربة تحديد خط الأساس لقدرة النموذج على الوصف الدلالي دون أي تكييف.

المقياس	القيمة
ROUGE-1	0.27
ROUGE-2	0.11
ROUGE-L	0.23
Semantic Similarity	0.36
Avg Generation Time	s / image1.2
Error Rate	%5

الجدول21 احتبار أداء SMOLVLM2-500M

كان النموذج يلتقط المعنى العام (وجود شخص/لون عام)، لكنه يفشل في الاتساق والتفاصيل الدقيقة.

المحاولة الثانية: Fine-tuning تقليدي كامل

هدفت هذه المحاولة إلى تحسين جودة الأوصاف النصية المولّدة من خلال تحديث جميع أوزان النموذج أثناء عملية التدريب، انطلاقًا من الافتراض القائل بأن التكييف الكامل قد يسمح للنموذج بمواءمة تمثيلاته الداخلية بصورة أدق مع خصائص بيانات المشاة المستهدفة.

وقد أظهرت النتائج المسجلة تحسّنًا طفيفًا في بعض المقاييس الكمية، مثل قيم ROUGE ومؤشر التشابه الدلالي (Semantic Similarity)، مقارنةً بالنموذج الأساسي غير المدرَّب.

إلا أن هذا التحسّن المحدود ترافق مع تكلفة حسابية مرتفعة، تمثّلت في استهلاك كبير للذاكرة الرسومية، وبطء ملحوظ في زمن التدريب، فضلًا عن عدم استقرار واضح في منحنيات التعلم. كما لوحظت مؤشرات مبكرة على فقدان المعرفة العامة للنموذج (Catastrophic Forgetting)، حيث أصبح النموذج أكثر تخصّصًا على حساب قدرته العامة على التوصيف المتوازن.

وبناءً على ذلك، خلصت هذه التجربة إلى أن Fine-tuning التقليدي الكامل غير ملائم للنماذج خفيفة الوزن مثل SmolVLM2، إذ إن المكاسب الدلالية المحدودة التي يحققها لا تبرّر الكلفة الحسابية العالية والمخاطر المنهجية المصاحبة له. وعليه، تم استبعاد هذا المسار من الإعدادات المعتمدة، والاتجاه نحو استراتيجيات تكييف خفيفة أكثر استقرارًا وكفاءة.

على الرغم من أن Fine-tuning التقليدي الكامل أظهر تحسّنًا محدودًا في بعض المؤشرات الكمية، مثل ROUGE والتشابه الدلالي، إلا أن هذه المكاسب لم تكن متناسبة مع الكلفة الحسابية المرتفعة وعدم الاستقرار الذي رافق عملية التدريب. فقد كشفت التجربة أن تحديث جميع أوزان نموذج خفيف الوزن مثل SmolVLM2 يؤدي إلى إرهاق بنيوي للنموذج، ويزيد من احتمالية فقدان المعرفة العامة، بدلًا من تعزيز قدرته الدلالية بشكل متوازن.

وعليه، تم الانتقال إلى اعتماد Fine-tuning خفيف باستخدام تقنية LoRA، التي تقوم على حقن معاملات قابلة للتعلّم ضمن طبقات محددة، مع تجميد الأوزان الأساسية للنموذج. وقد مكّن هذا التحول المنهجي من تحقيق توازن فعّال بين الاستقرار الحسابي والتحسّن الدلالي، كما سيتم توضيحه في نتائج المحاولة الثالثة.

المحاولة الثالثة: Fine-tuning خفيف باستخدام LoRA

هدفت هذه المحاولة إلى تحقيق تكييف دلالي فعّال لنموذج SmolVLM2 مع الحفاظ على استقراره البنيوي ومعرفته العامة، وذلك من خلال اعتماد استراتيجية Fine-tuning خفيف باستخدام تقنية LoRA (Low-Rank Adaptation). تقوم هذه الاستراتيجية على تجميد الأوزان الأساسية للنموذج، مع إدخال عدد محدود من المعاملات القابلة للتعلّم ضمن طبقات مختارة، ولا سيما طبقات الانتباه والإسقاط، بما يسمح بتوجيه النموذج نحو المهمة المستهدفة دون إعادة تشكيل فضائه التمثيلي بالكامل.

أظهرت النتائج الكمية تحسّنًا واضحًا مقارنةً بكل من النموذج الأساسي ومحاولة fine-tuning التقليدي الكامل. فقد ارتفعت قيمة ROUGE-1 من متوسط يقارب 0.28 في حالة النموذج الأساسي إلى نحو 0.33–0.35 بعد تطبيق LoRA، كما تحسّنت قيمة ROUGE-L من حوالي 0.24 إلى ما يقارب 0.29–0.31. وعلى المستوى الدلالي، سُجّل ارتفاع ملحوظ في مؤشر Semantic Similarity من نطاق 0.36–0.38 إلى نطاق أعلى بلغ 0.41–0.43، ما يشير إلى تحسّن حقيقي في اتساق الأوصاف المولّدة مع المعنى المرجعي، وليس مجرد تقارب لغوي سطحي.

من الناحية الحسابية، تحقق هذا التحسّن مع استقرار تدريبي مرتفع وانخفاض ملموس في استهلاك الذاكرة الرسومية مقارنةً بالتكييف الكامل، دون ظهور مؤشرات على فقدان المعرفة العامة أو تذبذب حاد في منحنيات التعلم. ويُبرز هذا السلوك أن التكييف الجزئي الموجَّه يمكّن النموذج من الاستفادة من معرفته المسبقة مع إدخال تعديلات دلالية كافية لمواءمة بيانات المشاة.

وبناءً على هذه النتائج، تبيّن أن Fine-tuning باستخدام LoRA يمثّل نقطة توازن منهجية بين جودة التمثيل الدلالي والكلفة الحسابية، وهو ما جعله الأساس الذي بُنيت عليه المرحلة اللاحقة من التجارب، والمتمثلة في دمج تحسينات Unsloth بهدف تعزيز الكفاءة العملية دون التأثير السلبي على الأداء الدلالي.

ولكن بقيت مخرجات النموذج تظهر أخطاءاً في ألوان الملابس (أخطأ النموذج بين ألوان مثل الكحلي والأبيض في بعض التجارب) الأمر الذي أوصلنا إلى قناعة أن النموذج أصغر من يستطيع استبدال النماذج الكبيرة المستخدمة في الأدبيات.

4.6.2. التحول من توصيف لغوي صحيح إلى تمثيل دلالي تمييزي

قمنا بعد ذلك بتجربة اعتماد النموذج اللغوي smolvlm2 في مقاربة شبيهة بالأدبيات لتوليد أشعة (عددية) دلالية للتمييز وتوصيف الصور بصيغة متوائمة مع مسألة إعادة التعرف، أي أن نجعل SmolVLM2 يكتب *وصفًا نصيًا صحيحًا* (Captioning) ثم نقيس ROUGE/BLEU، المشروع هنا يحاول استخدام الـ VLM كـ مولّد تمثيل دلالي (Semantic representation) يصلح لـ Re-ID:

- الـ VLM ‘‘يفهم’’ الصورة (ملابس/ألوان/حقيبة/عمر...)
- ثم نأخذ منه تمثيلًا عدديًا (vector/embedding) أو ‘‘توكن دلالي’’
- ثم ندرب رأس ReID (تصنيف + Triplet) ليجعل هذا التمثيل تمييزيا للهوية عبر الكاميرات

وهذا منسجم تمامًا مع فلسفة ورقة LVLM-ReID: الورقة تقول صراحة إن الاعتماد على captioning أو image-text matching قد لا يتوافق مع هدف ReID، ويقترحون بدل ذلك ‘‘توكن دلالي واحد’’ يُحسّن التمييز للهوية بدون حاجة لتعليقات نصية.

توليد التوكن الدلالي للمشاة (PSTG): المبدأ، البنية، ودور الإضافات المدرَّبة

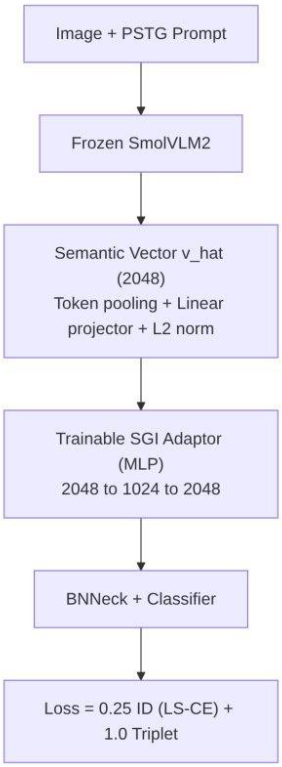
يعتمد مكوّن PSTG (Pedestrian Semantic Token Generation) على مبدأ استخدام النموذج اللغوي—البصري لاستخلاص تمثيل دلالي مضغوط يركّز على السمات الظاهرية العامة للمشاة، مثل نمط الملابس والألوان والملحقات، دون السعي إلى إنتاج توصيف لغوي كامل أو قابل للاستخدام البشري. ويُنظر إلى هذا التوكن الدلالي بوصفه تمثيلًا كامنًا وسيطًا يُفترض أن يحمل معلومات ذات صلة بالتمييز الهوياتي، يمكن توظيفها لاحقًا داخل بنية Re-ID.

من الناحية البنيوية، لا يتم في PSTG تحديث أوزان النموذج اللغوي—البصري نفسه، بل يُستخدم النموذج بوصفه مولِّدًا ثابتًا للإشارة الدلالية. ويتم استخراج التمثيل الدلالي (vhat) من مخرجات داخلية للنموذج (Hidden Representation)، وليس من النص النهائي المولَّد، وذلك لتجنّب الاعتماد على الصياغة اللغوية السطحية والتركيز بدلًا من ذلك على الفضاء الدلالي الكامن. هذا الاختيار يعكس التحوّل المنهجي في المشروع من تقييم ''صحة الوصف'' إلى تقييم ''قابلية التمثيل للتمييز''.

ولتكييف هذا التمثيل الدلالي مع متطلبات إعادة التعرف، أُضيفت وحدات قابلة للتدريب فوق vhat، تمثّلت في مكيّف تمثيلي (Adaptor) ورأس تمييزي خاص بـ Re-ID. يقوم المكيّف بتحويل التمثيل الدلالي الخام إلى فضاء أقل تشويشًا وأكثر توافقًا مع خسائر التدريب، دون افتراض أن التمثيل الأصلي يحمل بنية هوياتية كاملة. أما الرأس التمييزي، فيتكون من طبقة تطبيع (BN-like) تليها طبقة تصنيف، ويهدف إلى فرض فصل صريح بين الهويات أثناء التدريب.

تُفسَّر معاملات التدريب المرتبطة بهذا الإعداد بوصفها آليات تنظيم لا توليد معلومات جديدة؛ فمعامل خسارة التصنيف يحدّد مدى الضغط المفروض على التمثيل للفصل بين الهويات، بينما يضبط معامل خسارة الثلاثيات درجة إعادة تشكيل العلاقات الهندسية داخل الفضاء الكامن. غير أن النتائج العددية أظهرت أن هذه المعاملات — مهما بلغت دقتها — لا تستطيع تعويض غياب الإشارة الهوياتية في التمثيل الدلالي الأساسي، وهو ما انعكس في القيم المنخفضة جدًا لـ Rank-1 وmAP عند استخدام PSTG بصورة مستقلة.

وبناءً على ذلك، يتضح أن دور PSTG في هذا المشروع لم يكن فاشلًا من حيث المبدأ، بل كاشفًا لحدوده البنيوية؛ إذ أثبتت التجارب أن التوكن الدلالي، عند توليده واستخدامه بمعزل عن التمثيل البصري الأصلي، لا يشكّل أساسًا كافيًا لإعادة التعرف على الأشخاص. وتنسجم هذه النتيجة مع الفلسفة التي تتبناها الأطر الهجينة الحديثة، حيث يُفترض أن يعمل PSTG بوصفه آلية توجيه دلالي داخل مسار بصري تمييزي، لا كمصدر وحيد للتمثيل.



الشكل 32 النموذج اللغوي SMOLVLM2 + PSTG

قيم الأداء على مجموعة Market1501

اعتمد التقييم على البروتوكول القياسي لمجموعة Market1501، حيث جرى استخراج الأشعة الممثِّلة لصور الاستعلام والمعرض وتطبيعها باستخدام L2، ثم قياس التشابه بينها عبر المسافة الكوسينية. واتباعًا للإجراء المعتمد في المجموعة، استُبعدت صور المعرض التي تشترك مع صورة الاستعلام في كلٍّ من معرّف الشخص ومعرّف الكاميرا، لتجنّب المطابقة داخل نفس الكاميرا. وبناءً على مصفوفة المسافات الناتجة، حُسب كلٌّ من Rank-1 ضمن منحنى المطابقة التراكمية ومتوسط الدقة (mAP)، بما يوفّر تقييمًا دقيقًا لقدرة النموذج على إعادة التعرف على الأشخاص عبر الكاميرات المختلفة.

خسائر التدريب: المبررات والنتائج

تم اعتماد خسارتين تكميليتين أثناء التدريب: خسارة التصنيف وخسارة الثلاثيات. تفرض خسارة التصنيف فصلًا صريحًا بين الهويات المختلفة، في حين تنظّم خسارة الثلاثيات البنية الهندسية للفضاء التمثيلي عبر تقليل المسافة بين صور الهوية نفسها وزيادتها بين الهويات المختلفة. وقد ساهم الجمع بين الخسارتين في تحسين الاستقرار النظري للتعلم، إلا أن النتائج أظهرت أن

فعاليتهما تبقى مشروطة بجودة الأشعة الدلالية المدخلة؛ إذ إن ضعف التمثيل الأساسي ينعكس مباشرةً على قيم Rank-1 وmAP، مهما كانت صيغة الخسارة معتمدة نظريًا.

ضعف التمييزية في التمثيل الدلالي: دليل عددي مباشر

عند الاعتماد على الأشعة الدلالية المستخرجة مسبقًا (vhat) كمدخل وحيد لرأس ReID، أظهرت نتائج التقييم على مجموعة Market1501 انهيارًا شبه كامل في الأداء التمييزي. فقد سُجِّلت قيم منخفضة جدًا لكلٍّ من Rank-1 وmAP، ما يدل على أن التمثيل الدلالي — رغم احتوائه على معلومات وصفية عامة — لم يكن قادرًا على فصل الهويات المختلفة داخل فضاء الاسترجاع.

نتائج التقييم (Market1501)

تشير القيم شبه الصفرية لـ Rank-1 (**0.0009**) وmAP (**0.0053**) إلى أن ترتيب صور المعرض بالنسبة لصور الاستعلام كان قريبًا من العشوائي، وهو ما يعني عمليًا أن الأشعة الدلالية المستخرجة لم تحمل معلومات هوياتية قابلة للاستغلال، حتى بعد إدخال خسائر تصنيف وثلاثيات ورأس تمييزي.

وتؤكد هذه النتائج أن الخسائر المستخدمة لم تكن قادرة على تعويض غياب الإشارة الهوياتية في التمثيل الأساسي، وأن المشكلة لا تكمن في آلية التقييم أو صيغة الخسارة، بل في طبيعة التمثيل الدلالي نفسه عند استخدامه بصورة مستقلة دون دمجه مع تمثيل بصري تمييزي أو آلية تفاعل متعددة الوسائط.

وهذا منطقي بالمقارنة مع إطار عمل lvlm-reid الذي كان فيه دور التمثيل الدلالي توجيهياً

بالمقارنة مع الأدبيات لا يتعامل LVLM-ReID مع الدلالة بوصفها بديلًا عن التمثيل البصري، بل بوصفها إشارة مكمِّلة موجِّهة تُدمَج ضمن عملية الاستخراج نفسها. إذ يعتمد هذا الإطار على توليد توكن دلالي موجَّه بالتعليمات، ثم إدخاله في وحدة تفاعل تسمح له بالتأثير المباشر على التوكنات البصرية والعكس، قبل الوصول إلى التمثيل النهائي المستخدم في إعادة التعرف. وبهذا المعنى، تُستثمر الدلالة لتعزيز التمييز الهوياتي داخل الفضاء البصري، لا لاستخلاص تمثيل مستقل يُفترض أن يكون تمييزيًا بطبيعته. وهذا يفسر تماماً بالإضافة إلى صغر النموذج الذي نتعامل معه smolvlm2:500M.

4.7. دمج الخصائص Fusion Embeddings

4.7.1. دمج النماذج في أنظمة إعادة التعرّف

في أنظمة إعادة التعرّف على الأشخاص، غالبًا ما يتم الاعتماد على مصادر تمثيل مختلفة للخصائص، مثل خصائص المظهر المستخرجة من الصور، وخصائص الحركة أو البنية الحركية المستخرجة من تسلسل الإطارات (Tracklets). يهدف دمج هذه التمثيلات إلى الاستفادة من التكامل المعلوماتي بين مصادر متعددة، بحيث يعوّض كل تمثيل نقاط ضعف الآخر.

تنقسم استراتيجيات الدمج عمومًا إلى فئتين رئيسيتين:

- **الدمج المبكر (Early Fusion)**
- **الدمج المتأخر (Late Fusion)**

الدمج المبكر مقابل الدمج المتأخر

- **الدمج المبكر (Early Fusion)**

يعتمد الدمج المبكر على دمج مصادر البيانات أو الميزات الخام في مراحل مبكرة من النموذج، قبل أو أثناء عملية التعلم. فعلى سبيل المثال، يمكن دمج صور متعددة أو خرائط خصائص مختلفة وإدخالها إلى شبكة واحدة مشتركة.

ورغم أن هذا الأسلوب قد يتيح للنموذج تعلم علاقات مشتركة بين المصادر المختلفة، إلا أنه يعاني من عدة قيود جوهرية في سياق هذه الدراسة:

- اختلاف طبيعة البيانات (صورة مفردة مقابل تسلسل زمني).
- اختلاف الأبعاد الإحصائية والدلالية للخصائص.
- الحاجة إلى إعادة تدريب النموذج بالكامل عند إضافة مصدر جديد للخصائص.

- **الدمج المتأخر (Late Fusion)**

في المقابل، يعتمد الدمج المتأخر على استخراج متجه تمثيلي مستقل من كل نموذج على حدة، ثم دمج هذه المتجهات في مرحلة لاحقة. يتميز هذا الأسلوب بعدة مزايا جعلته الخيار الأنسب في هذا العمل:

- استقلالية النماذج: كل نموذج يُدرَّب ويُقيَّم بشكل منفصل.
- مرونة عالية في إضافة أو إزالة مصادر خصائص.

- وضوح تحليلي يسمح بدراسة مساهمة كل تمثيل على حدة.
- تقليل التداخل السلبي بين تمثيلات غير متجانسة.

بناءً على ذلك، تم اعتماد الدمج المتأخر كإطار منهجي لدمج خصائص المظهر وخصائص الحركة.

فكرة دمج الأشعة بالـ Concatenation

يعتمد الدمج المتأخر في هذا العمل على عملية الربط التسلسلي (Concatenation) لأشعة الخصائص.

الشكل33 شعاع دمج مركب من عدة أشعة ناتجة عن نماذج مختلفة وبأبعاد مختلفة

وتقوم هذه العملية على مبدأ بسيط وفعّال فإذا كان لدينا شعاعان تمثيليان مستقلان:

- شعاع خصائص المظهر
- شعاع خصائص الحركة (Kinematic)

فإن دمجهما يتم عبر وضعهما جنبًا إلى جنب لتكوين متجه واحد ذي بعد أعلى، بحيث يحتفظ كل جزء من المتجه المدمج بمعلوماته الأصلية دون إسقاط أو تلخيص مبكر.

عند عملية الدمج، يتم تطبيق تطبيع L2 على المتجه الناتج لضمان استقرار المقارنة باستخدام مقاييس المسافة، ومنع سيطرة أحد التمثيلين بسبب اختلاف السعة العددية.

دمج نماذج الصورة المفردة مع نماذج الـ Tracklet

تواجه عملية دمج خصائص مستخرجة من صورة واحدة مع خصائص مستخرجة من تسلسل إطارات تحديًا بنيويًا، يتمثل في عدم التوافق المباشر بين تمثيل ثابت وتمثيل زمني.

لمعالجة هذا الإشكال، تم اعتماد استراتيجية تقوم على اختيار ممثل (Representative) واحد عن كل Tracklet، بحيث يتم اختزال التسلسل الزمني إلى متجه واحد قابل للدمج مع خصائص الصورة المفردة.

قمنا بتجربة على نموذج المظهر الهجين الخاص بنا مع مجموعة بيانات فيديوية هي MARS لتمثيل كل tracklets بشعاع خاص بإطار واحد فقط بدلا من شعاع لكل إطار بهدف أمكانية توليد شعاع ناتج عن دمج نموذجين + single image tracklet في شعاع واحد

قمنا بمحاولة اختيار الإطار الذي سيمثل أطر ال tracklet بمحصلتها وكانت النتائج

Rank-20	Rank-10	Rank-5	Rank-1	mAP	Strategy (per tracklet)
%72.22	%65.53	%59.26	%41.45	%39.28	Mean of all frames (our original tracklet emb)
%47.29	%39.74	%34.19	%19.52	%18.24	1st frame only
%48.72	%41.60	%34.62	%20.37	%18.94	2nd frame only
%50.00	%44.44	%37.18	%20.66	%19.86	3rd frame only
%52.85	%45.30	%38.75	%23.36	%21.48	4th frame only
%51.85	%45.30	%38.18	%22.93	%21.66	5th frame only

الجدول22 جدول نتائج تجربة خيارات الإطار الممثل للسلسلة مع متوسطها الحسابي

ومنه تبين لنا أنه دوما من الأفضل بدلاً من أن نختار إطاراً بعينه من مجموعة أطر عنصر المرور tracklet فإن المتوسط المحسوب منها جميعاً هو ممثل أفضل من أي منها، كما أن هذا الخيار يحل إشكالية كون عناصر المرور متغيرة في عدد الأطر وكون المشيات تجري بسرعات مختلفة فليس من المنطقي أن يكون إطار واحد ثابت الترتيب هو الممثل الأفضل.

تحليل نتائج الدمج

تُظهر النتائج التجريبية ما يلي:

- نموذج المظهر فقط يحقق أداءً مرتفعًا نسبيًا، ما يؤكد قوة الخصائص البصرية الثابتة.
- نموذج KinematicGNN بمفرده يحقق أداءً ضعيفًا للغاية، ما يشير إلى أن الخصائص الحركية وحدها غير كافية للتمييز في هذا الإعداد.
- الدمج عبر Concatenation أدى إلى تحسن واضح مقارنة بالنموذج الحركي وحده، لكنه لم يتجاوز أداء نموذج المظهر فقط.

يمكن تفسير هذا السلوك بأن:

- الخصائص الحركية تضيف معلومات مكملة، لكنها لا تزال محدودة التمييز مقارنة بالمظهر.
- الدمج المتأخر حافظ على استقرار النظام ومنع تدهور الأداء، لكنه كشف أيضًا أن جودة التمثيل الحركي عامل حاسم في فعالية الدمج.

تجربة دمج نموذجي المظهر والقناة الحركية

شملت التجربة ثلاث حالات رئيسية للمقارنة:

1. نموذج المظهر فقط (Appearance-only)

في هذه الحالة، تم تمثيل كل عنصر مرور (Tracklet) باستخدام متجه مظهر واحد ناتج عن حساب متوسط متجهات جميع الإطارات ضمن الـ Tracklet، بهدف الحصول على تمثيل ثابت يعكس الخصائص البصرية العامة للشخص عبر الزمن.

2. نموذج الخصائص الحركية فقط (KinematicGNN-only)

في هذه الحالة، تم الاعتماد على متجه الخصائص الحركية المستخرج من نموذج KinematicGNN كمصدر وحيد للتمثيل، دون دمجه مع أي معلومات مظهرية.

3. التمثيل المدمج (Fusion)

في هذه الحالة، تم دمج متجه المظهر مع متجه الخصائص الحركية باستخدام عملية الربط التسلسلي (Concatenation)، تلاها تطبيق تطبيع من نوع L2 على المتجه الناتج. وقد تم الحرص على توحيد آلية التمثيل بحيث يُمثَّل كل عنصر مرور بمتجه واحد فقط، مما يضمن عدالة المقارنة مع الحالتين السابقتين.

```
>>> Appearance-only (mean over frames)
mAP   : 39.28%
Rank-1: 41.45%
Rank-5: 59.26%
Rank-10: 65.53%
Rank-20: 72.22%
>>> KinematicGNN-only
mAP   : 0.62%
Rank-1: 0.28%
Rank-5: 1.28%
Rank-10: 1.71%
Rank-20: 3.99%
>>> Fusion (appearance || kinematic, concat + L2)
fused query embeddings: (702, 1152), gallery embeddings: (2636, 1152)
mAP   : 19.07%
Rank-1: 24.64%
Rank-5: 34.19%
Rank-10: 37.04%
Rank-20: 40.60%
```

الجدول23 جدول نتائج تجارب تطبيق الدمج بين نموذجين

نلاحظ ما يلي:

- النموذج المدرب على صورة واحدة وبعد تطبيقه على معطيات مهيكلة بشكل tracklets واختيار ممثل قد أبدى انخفاضاً واضحاً في الأداء

- انخفاض واضح في أداء النموذج kinematicGNN لدى تطبيقه على مجموعة بيانات فيديوية ذات حالات ضجيج على مدخلات بيانات وضعيات مستخرجة باستخدام yolo8-pose كما كان متوقعاً لأننا نتعامل مع مجموعة بيانات مختلفة في النمط إضافة إلى أن MARS هي مجموعة بيانات مشهورة باحتوائها حالات ضجيج وحالات حجب كبيرة وعديدة (تسبب هذه الحالات عدم التعرف على الوضعية بتاتاً أو تقديرات خاطئة لإحداثيات عقد الاطراف المحجوبة غالباً ما تكون الأقدام)
- أدى نموذج الدمج أداء وسيطاً متناسباً مع جودة النتائج لكلا النموذجين وهذا موافق للتوقعات كون نموذج الدمج المعتمد هو إلصاق للأشعة ببعضها بشكل متتالي وتوليد شعاع واحد كبير.

4.8.خاتمة

يوضح هذا الفصل حدود على أهمية دمج الأنظمة المتعتمدة على خصائص مختلفة ويؤكد على دراسة آليات دمج خصائص أكثر تعقيداً بهدف الحصول على أفضل ما يمكن من كل نموذج ، كما يوضح الحدود التي تحكم الاعتماد على كل خاصية من الخصائص وحدود العمل في تشغيل النماذج ضمن بيئات العمل الحقيقية على الحافة ضمن أنظمة مضمنة وعلى المخدمات ويشكل مرجعاً جيداً لبدء بناء نظام إعادة تعرف بشكل عملي وهندسي.

الفصل الخامس: الخاتمة والآفاق المستقبلية

5.1. الخاتمة

في ختام هذه الدراسة، تناولنا مسألة إعادة التعرّف على الأشخاص بوصفها مشكلة متعددة الأبعاد، لا يمكن مقاربتها من زاوية واحدة أو عبر نوع واحد من الخصائص. فقد أظهر التحليل المنهجي أن الخصائص الزمنية المرتبطة بالحركة، مثل تغيّر بنية الجسم عبر الإطارات وإيقاع المشية، تقدّم معلومات تمييزية عميقة تعكس هوية الفرد بطريقة أقل تأثرًا بتغيرات المظهر، لكنها في المقابل تتطلب شروطًا صارمة تتعلق بجودة التسلسل الزمني واستمراريته. في المقابل، أثبتت الخصائص المعتمدة على المظهر فعاليتها في البيئات العملية واسعة النطاق بفضل بساطتها وكفاءتها الحسابية، إلا أنها تظل حساسة لتغيرات الإضاءة، زاوية الرؤية، والملابس.

كما بيّنت الدراسة أن الصفات الدلالية المستخلصة عبر النماذج اللغوية—البصرية تمثل اتجاهًا واعدًا لتجاوز بعض القيود التقليدية، من خلال توفير تمثيل عالي المستوى يقرّب بين الإدراك البشري والتمثيل الحاسوبي. غير أن هذا المسار ما يزال يواجه تحديات تتعلق بالاتساق الدلالي، الكلفة الحسابية، وصعوبة دمجه المباشر في أنظمة الاسترجاع التقليدية دون تصميم معماري مدروس.

انطلاقًا من ذلك، خلصت هذه الدراسة إلى أن الحل الأكثر واقعية ومرونة لا يكمن في تفضيل نوع واحد من الخصائص، بل في تبنّي مقاربات هجينة تستثمر التكامل بين المظهر، الحركة، والزمن، والدلالة، مع مراعاة شروط التشغيل الفعلية لكل نظام. وقد أظهرت التجارب والتحليلات أن استراتيجيات الدمج المتأخر تمثل خيارًا عمليًا لتحقيق هذا التوازن، إذ تسمح بالاستفادة من نقاط القوة لكل تمثيل مع الحد من تداخل القيود فيما بينها.

أخيرًا، تبرز نتائج هذه الدراسة الحاجة إلى إعادة التفكير في تصميم مجموعات البيانات وبروتوكولات التقييم، بما يعكس بشكل أدق سيناريوهات المراقبة الواقعية، ويفتح المجال أمام نماذج أكثر تكيفًا وقابلية للتعميم. وعليه، تشكّل هذه الأطروحة خطوة نحو فهم أعمق لمشكلة إعادة التعرّف، وتمهّد الطريق لأبحاث مستقبلية تسعى إلى بناء أنظمة أكثر موثوقية، تفسيرية، وقربًا من متطلبات التطبيقات الحقيقية.

5.2. الآفاق المستقبلية (Research Directions)

تفتح هذه الدراسة عددًا من المسارات البحثية التي يمكن أن تسهم في تعميق فهم مسألة إعادة التعرّف وتجاوز القيود الحالية للنماذج المعتمدة. أول هذه المسارات يتمثل في استكشاف نماذج دمج أكثر تعقيدًا تتجاوز الدمج الشعاعي التقليدي، وذلك عبر توظيف التمثيل البياني للجسم البشري واستخدام عقد مرجعية (Anchors) تعمل كمرتكزات لربط الخصائص المميزة المستخلصة من النماذج المختلفة، بما يسمح بنمذجة العلاقات البنيوية والدلالية بين هذه الخصائص بصورة أكثر صراحة.

كما يبرز اتجاه بحثي واعد يتمثل في متابعة تطوير أطر شبيهة بـ LVLM-ReID ولكن بالاعتماد على نماذج لغوية–بصرية صغيرة الحجم، بما يتيح دراسة التوازن بين القدرة التمثيلية والكلفة الحسابية، ويفتح المجال أمام استخدام هذه المقاربات في سيناريوهات ذات موارد محدودة.

ومن الآفاق المهمة كذلك دمج النموذج اللغوي مع نموذج رديف (Auxiliary Model) لمعالجة الأخطاء الدلالية الشائعة، مثل الالتباس في توصيف الألوان أو إسنادها الخاطئ لأجزاء الجسم. ويمكن تحقيق ذلك عبر ربط المخرجات الدلالية بالتمثيل البياني للوضعيات البشرية، بما يسمح بمواءمة الوصف اللغوي مع البنية الهندسية للجسم.

إضافة إلى ذلك، تشكّل مشكلة اختلاف الكاميرات وانحيازاتها محورًا بحثيًا مفتوحًا، حيث يمكن التعمق في تطوير استراتيجيات دمج تشمل أشعة (Embeddings) خاصة بالكاميرات، تُحقن ضمن التمثيل المكمم للشعاع الناتج عن النموذج، بهدف تمكين النظام من استيعاب الخصائص السياقية للكاميرا والتخفيف من تأثير الفجوة بين المجالات (Domain Gap).

وأخيرًا، تبرز الحاجة إلى إجراء أبحاث معمقة لرفع أداء نماذج المشية عند تطبيقها على مجموعات بيانات فيديوية حقيقية، حيث تتداخل تحديات الزمن، الضجيج، وعدم اكتمال التسلسلات، وهو ما يتطلب نماذج أكثر قدرة على التكيّف مع الطبيعة غير المثالية للبيانات الواقعية.

الاتجاه في البحث ضمن مجال ذكاء الوكلاء Agentic AI وبدلا اعتماد استراتيجية دمج ثابتة، يمكن لوكلاء ذكيين أن يقوموا باختيار استراتيجية الدمج المناسبة ديناميكيًا وفقًا لسياق المشهد، جودة البيانات، أو حالة النظام. فعلى سبيل المثال، في حال ضعف جودة المظهر، يمكن للوكيل تعزيز وزن المشية أو السمات الزمنية، بينما يُفضَّل المظهر في الحالات الساكنة. يمثل هذا الاتجاه خطوة نحو أنظمة إعادة تعرّف واعية بالسياق (Context-Aware ReID).

5.3. التوصيات الهندسية (Engineering Recommendations)

على المستوى التطبيقي، توصي هذه الدراسة بإجراء تجارب تشغيلية أكثر تنوعًا على بيئات حوسبة حافّية حقيقية، بما يسمح بتقييم النماذج ليس فقط من حيث الدقة، بل أيضًا من حيث الاستقرار، زمن الاستجابة، واستهلاك الموارد، وهي عوامل حاسمة في الأنظمة العاملة ميدانيًا.

كما يُنصح باستكشاف بنى تخزين معطيات متقدمة قادرة على تمثيل البعدين الزمني والمكاني بكفاءة، مثل قواعد المعطيات البيانية والزمنية، بهدف تحسين أداء عمليات الاسترجاع المرتبطة بإعادة التعرّف، لا سيما عند التعامل مع تسلسلات طويلة وسيناريوهات متعددة الكاميرات.

وفي السياق ذاته، تمثل تصميم بنية نظام ذاكرة خبيئة (Cache System) متوافقة مع خصائص نظام إعادة التعرّف توصية هندسية محورية، إذ يمكن عبرها تقليل زمن الاستجابة وتحسين قابلية التوسع، من خلال تخزين الأشعة والنتائج الوسيطة الأكثر استخدامًا بطريقة مدروسة تتماشى مع طبيعة الاستعلامات في أنظمة المراقبة.

ختامًا، يهدف هذا الفصل إلى توفير خارطة طريق مزدوجة تجمع بين التطوير البحثي طويل الأمد والاعتبارات الهندسية العملية، بما يمكّن الباحثين والمهندسين على حد سواء من البناء على نتائج هذه الدراسة وتوجيه جهودهم نحو أنظمة إعادة تعرّف أكثر كفاءة، مرونة، وقابلية للاستخدام في البيئات الواقعية.

المراجع

[1] K. Sun, B. Xiao, D. Liu, and J. Wang, “Deep High-Resolution Representation Learning for Human Pose Estimation,” 2019, *arXiv*. doi: 10.48550/ARXIV.1902.09212.

[2] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, “A Comprehensive Survey on Graph Neural Networks,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 4–24, Jan. 2021, doi: 10.1109/TNNLS.2020.2978386.

[3] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, “Graph Attention Networks,” Feb. 04, 2018, *arXiv*: arXiv:1710.10903. doi: 10.48550/arXiv.1710.10903.

[4] A. Vaswani *et al.*, “Attention Is All You Need,” Aug. 02, 2023, *arXiv*: arXiv:1706.03762. doi: 10.48550/arXiv.1706.03762.

[5] A. Dosovitskiy *et al.*, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” June 03, 2021, *arXiv*: arXiv:2010.11929. doi: 10.48550/arXiv.2010.11929.

[6] Z. Liu *et al.*, “Swin Transformer: Hierarchical Vision Transformer using Shifted Windows,” Aug. 17, 2021, *arXiv*: arXiv:2103.14030. doi: 10.48550/arXiv.2103.14030.

[7] E. J. Hu *et al.*, “LoRA: Low-Rank Adaptation of Large Language Models,” Oct. 16, 2021, *arXiv*: arXiv:2106.09685. doi: 10.48550/arXiv.2106.09685.

[8] T.-Y. Lin *et al.*, “Microsoft COCO: Common Objects in Context,” Feb. 21, 2015, *arXiv*: arXiv:1405.0312. doi: 10.48550/arXiv.1405.0312.

[9] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, “Scalable Person Re-identification: A Benchmark,” in *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile: IEEE, Dec. 2015, pp. 1116–1124. doi: 10.1109/ICCV.2015.133.

[10] M. Gou, Z. Zhang, A. Li, and L. Zheng, “DukeMTMC4ReID: A Large-Scale Multi-Camera Person Re-Identification Dataset,” in

Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2017.

[11] Y. Wu, Y. Lin, X. Dong, Y. Yan, W. Ouyang, and Y. Yang, "Exploit the Unknown Gradually: One-Shot Video-Based Person Re-Identification by Stepwise Learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. doi: 10.1109/CVPR.2018.00543.

[12] L. Zheng *et al.*, "MARS: A Video Benchmark for Large-Scale Person Re-identification," in *European Conference on Computer Vision (ECCV)*, 2016. doi: 10.1007/978-3-319-46484-8_52.

[13] Z. Ding, Y. Yang, H. Cheng, M. Shao, and Y. Fu, "Semantically Self-Aligned Network for Text-to-Image Part-Aware Person Re-Identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, pp. 1419–1427.

[14] W. Li, R. Zhao, T. Xiao, and X. Wang, "DeepReID: Deep Filter Pairing Neural Network for Person Re-identification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 152–159. doi: 10.1109/CVPR.2014.27.

[15] S. Yu, D. Tan, and T. Tan, "A Framework for Evaluating the Effect of View Angle, Clothing and Carrying Condition on Gait Recognition," in *Proceedings of the IEEE International Conference on Pattern Recognition (ICPR)*, 2006.

[16] N. Takemura, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "Multi-view Large Population Gait Dataset and Its Performance Evaluation for Cross-View Gait Recognition," *IPSJ Trans. Comput. Vis. Appl.*, vol. 10, no. 1, 2018, doi: 10.1186/s41074-018-0035-3.

[17] Y. Zhu, Y. Chen, C. Liang, S. Wang, and Q. Tian, "GREW: A Large-Scale Gait Recognition Dataset in the Wild," *ArXiv Prepr. ArXiv210502516*, 2021, [Online]. Available: https://arxiv.org/abs/2105.02516

[18] R. Liao, S. Yu, W. An, and Y. Huang, "A model-based gait recognition method with body pose and human prior knowledge,"

Pattern Recognit., vol. 98, p. 107069, Feb. 2020, doi: 10.1016/j.patcog.2019.107069.

[19] T. Teepe, A. Khan, J. Gilg, F. Herzog, S. Hörmann, and G. Rigoll, “GaitGraph: Graph Convolutional Network for Skeleton-Based Gait Recognition,” in *2021 IEEE International Conference on Image Processing (ICIP)*, Sept. 2021, pp. 2314–2318. doi: 10.1109/ICIP42928.2021.9506717.

[20] J. Wang, E. Bergeret, and I. Falih, “Skeleton-Based Action Recognition with Spatial-Structural Graph Convolution,” July 31, 2024, *arXiv*: arXiv:2407.21525. doi: 10.48550/arXiv.2407.21525.

[21] V. P. Dwivedi and X. Bresson, “A Generalization of Transformer Networks to Graphs,” Jan. 24, 2021, *arXiv*: arXiv:2012.09699. doi: 10.48550/arXiv.2012.09699.

[22] H. Rao and C. Miao, “Motif Guided Graph Transformer with Combinatorial Skeleton Prototype Learning for Skeleton-Based Person Re-Identification,” Feb. 02, 2025, *arXiv*: arXiv:2412.09044. doi: 10.48550/arXiv.2412.09044.

[23] H. Chao, Y. He, J. Zhang, and J. Feng, “GaitSet: Regarding Gait as a Set for Cross-View Gait Recognition,” Dec. 12, 2018, *arXiv*: arXiv:1811.06186. doi: 10.48550/arXiv.1811.06186.

[24] C. Fan *et al.*, “GaitPart: Temporal Part-Based Model for Gait Recognition,” in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA: IEEE, June 2020, pp. 14213–14221. doi: 10.1109/CVPR42600.2020.01423.

[25] C. Fan, S. Hou, Y. Huang, and S. Yu, “Exploring Deep Models for Practical Gait Recognition,” Jan. 10, 2024, *arXiv*: arXiv:2303.03301. doi: 10.48550/arXiv.2303.03301.

[26] C. Fan, J. Ma, D. Jin, C. Shen, and S. Yu, “SkeletonGait: Gait Recognition Using Skeleton Maps”.2025

[27] X. Zhang *et al.*, “TE-TransReID: Towards Efficient Person Re-Identification via Local Feature Embedding and Lightweight Transformer,” *Sensors*, vol. 25, no. 17, p. 5461, Sept. 2025, doi: 10.3390/s25175461.

[28] S. Li, L. Sun, and Q. Li, “CLIP-ReID: Exploiting Vision-Language Model for Image Re-Identification without Concrete Text Labels,” Jan. 01, 2023, *arXiv*: arXiv:2211.13977. doi: 10.48550/arXiv.2211.13977.

[29] Q. Wang, B. Li, and X. Xue, “When Large Vision-Language Models Meet Person Re-Identification,” Nov. 27, 2024, *arXiv*: arXiv:2411.18111. doi: 10.48550/arXiv.2411.18111.