UNIVERSITÉ DE YAOUNDÉ I
UNIVERSITY OF YAOUNDE I
****

DÉPARTEMENT D'INFORMATIQUE
DEPARTMENT OF COMPUTER SCIENCE

****

Optimisation de la détection multi-visages sur mobile par apprentissage profond pour un système de reconnaissance faciale hors ligne : application au marquage de présence en milieu académique

Mémoire présenté et soutenu par :

FOKOU Arnaud Cedric – 13Y180

Sous l a direction de :
Dr. AMINOU Halidou
Dr. Janvier NGNOULAYE

Année Académique :
2024 - 2025

♣ Table des matières ♣

Liste des abréviations **ii**

Table des Figures **iii**

Liste des tableaux **iv**

1 Introduction Générale **1**
- 1.1 Contexte général de l'étude . . . 1
- 1.2 Problématique . . . 2
- 1.3 Questions de recherche . . . 3
- 1.4 Objectifs de recherche . . . 3
- 1.5 Importance de l'étude . . . 4
- 1.6 Plan du mémoire . . . 4

2 Revue de la littérature **5**
- 2.1 Vision par Ordinateur et reconnaissance de Formes . . . 5
- 2.2 Systèmes Biométriques . . . 6
- 2.3 Analyse Faciale . . . 7
- 2.4 Détection et Suivi de Visages . . . 8
- 2.5 Apprentissage Automatique . . . 9
- 2.6 Architectures d'apprentissage profond . . . 9
- 2.7 Architecture single-stage . . . 10
- 2.8 Etat de l'art . . . 11

3 Méthodologie **20**
- 3.1 la collecte et le prétraitement des données . . . 20
- 3.2 Détection de Visages . . . 20
- 3.3 Identification de Visages . . . 20
- 3.4 Optimisation . . . 21
- 3.5 Évaluation . . . 21

4 Résultats et Discussions **22**
4.1 Résultats 22
4.1.1 Discussion 22
4.2 Conclusion 23
4.3 Perspectives 23

Conclusion Générale et Perspectives **23**

Bibliographie **23**

♣ Table des figures ♣

2.1 Performance vs Complexité des détecteurs de visages 15
2.2 Techniques d'analyse pour les systèmes biométriques 18

♣ Liste des tableaux ♣

2.1 Analyse comparative des approches de détection de visages triées par mAP, Paramètres et GFLOPs (ordre décroissant) 14
2.2 Analyse comparative des méthodes de détection de visage 16

Chapitre un

Introduction Générale

1.1 Contexte général de l'étude

La biométrie faciale, révolutionne les méthodes d'authentification dans de nombreux secteurs. Cette technologie trouve des applications diverses, de la sécurité aéroportuaire au déverrouillage des smartphones, en passant par la surveillance urbaine et le contrôle d'accès en entreprise. Dans le secteur éducatif en particulier, où l'optimisation des processus administratifs devient cruciale, la biométrie faciale offre des perspectives prometteuses pour la gestion automatisée des présences Raj et al. (2024)

L'optimisation des systèmes de reconnaissance faciale pour les appareils mobiles représente aujourd'hui un défi majeur dans le domaine de l'apprentissage profond, particulièrement en mode hors ligne. Dans les établissements d'enseignement, où le suivi des présences reste une tâche chronophage, les méthodes traditionnelles comme les appels nominaux et les feuilles d'émargement montrent leurs limites, étant non seulement chronophages mais aussi vulnérables à la fraude. La reconnaissance faciale sur mobile s'impose comme une solution pertinente face à d'autres alternatives : contrairement aux systèmes d'empreintes digitales qui nécessitent une authentification séquentielle, ou aux badges RFID facilement transférables, elle permet une identification simultanée et non falsifiable de multiples étudiants Brown (2021). De plus, en s'appuyant sur les smartphones déjà disponibles, cette approche évite tout investissement matériel supplémentaire tout en garantissant un fonctionnement hors ligne adapté aux contraintes de connectivité des établissements Bhat et al. (2020).

L'émergence des techniques d'apprentissage profond, en particulier les réseaux de neurones convolutifs (CNN), a transformé la reconnaissance faciale. Toutefois, leur déploiement sur appareils mobiles est confronté à des contraintes significatives : ressources matérielles limitées, nécessité d'un fonctionnement hors ligne, et variations environnementales en salle de classe. L'optimisation de ces modèles pour maintenir des performances

élevées sous ces contraintes constitue un défi technique majeur Alonso-Fernandez et al. (2024)

1.2 Problématique

Les systèmes de reconnaissance faciale basés sur l'apprentissage profond comprennent deux phases principales : la détection des visages dans l'image et la génération des enregistrements de caractéristiques (feature embeddings) Deng et al. (2023). Si ces systèmes atteignent aujourd'hui des performances remarquables sur des images de haute qualité, leur déploiement sur appareils mobiles pour le marquage de présence en milieu académique se heurte à la réalité des images capturées en classe : résolution limitée (typiquement 640x480 pixels ou moins après compression), qualité variable, et présence de multiples visages Khabarlak (2022).

Dans ce pipeline de traitement, la phase de détection multi-visages sur images de faible résolution constitue le défi majeur. En effet, alors que la génération d'embeddings peut être optimisée par un pré-enregistrement contrôlé des étudiants avec des images de bonne qualité, la détection doit traiter des images où chaque visage peut n'occuper que 20x20 à 32x32 pixels, avec une qualité dégradée par :

- La compression automatique des images (JPEG avec facteur de qualité souvent inférieur à 75
- Les conditions d'éclairage variables des salles de classe (mixte naturel/artificiel)
- Les mouvements lors de la capture créant du flou
- Les occlusions partielles entre étudiants

La nécessité d'un fonctionnement hors ligne, imposée par le contexte académique où la connexion internet n'est pas toujours fiable, ajoute une contrainte supplémentaire : tout le traitement doit être effectué localement sur le mobile avec des ressources limitées.

Dans ce contexte, la problématique centrale réside dans l'optimisation d'un système de détection multi-visages basé sur l'apprentissage profond pour traiter efficacement des images de faible résolution sur mobile en mode hors ligne. Cette optimisation doit surmonter plusieurs défis techniques :

- L'adaptation des architectures CNN pour détecter des visages de 20x20 à 32x32 pixels avec une précision supérieure à 90

- L'optimisation des modèles pour maintenir un temps de traitement inférieur à 500ms par image sur mobile
- La gestion des dégradations d'image avec un taux de faux négatifs inférieur à 5% sous différentes conditions d'éclairage
- Le maintien d'une empreinte mémoire inférieure à 100MB pour l'ensemble du système

1.3 Questions de recherche

Comment optimiser un système de détection multi-visages par apprentissage profond pour une exécution efficace sur mobile en mode hors ligne dans un contexte de marquage de présence académique ?

- Quel impact ont les différentes architectures CNN sur les performances de détection multi-visages en termes de précision et de ressources mobiles ?
- Comment améliorer la robustesse du système face aux variations d'éclairage et aux occlusions dans un environnement de classe ?
- Quelles techniques d'optimisation permettent de maintenir les performances en mode hors ligne sous contraintes mobiles ?

1.4 Objectifs de recherche

Développer un système optimisé de détection multi-visages par apprentissage profond pour le marquage automatique des présences sur mobile en mode hors ligne.

- Évaluer l'impact des différentes architectures CNN sur la performance de détection multi-visages pour smartphones, en considérant le compromis précision/ressources ;
- Implémenter des techniques d'amélioration de la robustesse du système face aux variations d'illumination et aux occlusions dans un environnement de classe ;
- Optimiser l'exécution du système en mode hors ligne à travers des techniques de compression et d'accélération adaptées aux contraintes mobiles.

1.5 Importance de l'étude

Cette recherche revêt une importance significative à plusieurs niveaux :

- Académique : Elle contribue à l'optimisation des modèles d'apprentissage profond pour les systèmes mobiles, proposant des solutions innovantes pour l'exécution hors ligne de tâches complexes.
- Technique : L'étude développe des techniques d'optimisation applicables à d'autres domaines nécessitant le déploiement de modèles d'apprentissage profond sur appareils mobiles.
- Pratique : Le système apporte une solution concrète aux établissements d'enseignement pour la gestion automatisée des présences, adaptée aux contraintes réelles du terrain.

1.6 Plan du mémoire

Cette etude comprendra les chapitres suivants :

- **Revue de la littérature (chapitre 2) :** ce chapitre passe en revue la littérature sur les systèmes de reconnaissance faciale, les réseaux de neurones convolutifs (CNN), les techniques d'optimisation des modèles, les défis de la reconnaissance faciale en temps réel sur mobile, et les applications de la reconnaissance faciale dans l'éducation.
- **Méthodologie (chapitre 3) :** ce chapitre comprend l'approche méthodologique, incluant la collecte et le prétraitement des données, l'architecture du système de reconnaissance faciale proposé, les techniques d'optimisation pour mobile, et le protocole expérimental avec les métriques d'évaluation.
- **Résultats et discussions (chapitre 4) :** les résultats de l'étude sont présentés dans ce chapitre, ainsi que des informations sur les performances du système de reconnaissance faciale en décrivant l'impact des différentes architectures CNN, les effets de la quantification, et l'influence des techniques d'optimisation.
- **Conclusion (chapitre 5) :** ce chapitre présente des recommandations pour l'optimisation de la détection des visages dans des systèmes de reconnaissance faciale destinés aux salles de classe, décrit les principaux résultats de l'étude et réaffirme les objectifs de la recherche.

Chapitre Deux

Revue de la littérature

La détection multi-visages sur plateforme mobile pour le marquage de présence académique représente un défi technique complexe, situé à l'intersection de plusieurs domaines de recherche. Cette revue examine l'évolution des approches et justifie les choix technologiques conduisant à notre solution.

2.1 Vision par Ordinateur et reconnaissance de Formes

La vision par ordinateur et la reconnaissance de formes se divisent en trois approches majeures pour l'analyse d'images numériques : l'analyse de scènes, la compréhension d'images et les systèmes biométriques. Chacune de ces approches propose une perspective différente pour le traitement de l'information visuelle.

L'analyse de scènes (Scene Analysis) se concentre sur la compréhension globale de l'environnement visuel. Cette approche segmente l'image en régions sémantiques et interprète les relations spatiales entre les objets Mohammed and Ralescu (2024). Dans un contexte de classe, elle permettrait de comprendre l'agencement général de la salle et la disposition des étudiants. Cependant, cette approche globale, bien que riche en informations contextuelles, s'avère trop générale et computationnellement coûteuse pour notre objectif spécifique de détection de visages. Les modèles d'analyse de scènes nécessitent généralement plusieurs gigaoctets de mémoire et des temps de traitement dépassant la seconde, les rendant inadaptés aux contraintes mobiles.

La compréhension d'images (Image Understanding) vise à extraire une interprétation sémantique complète du contenu visuel. Cette approche identifie non seulement les objets présents mais aussi leurs attributs et leurs interactions Astolfi et al. (2021). Appliquée à notre contexte, elle fournirait des informations sur l'attention des étudiants, leur posture, ou même leur niveau d'engagement. Bien que ces informations soient potentiellement utiles, cette approche généraliste nécessite des modèles complexes (>500MB) et un

temps de traitement significatif (>2s par image), dépassant largement nos contraintes de ressources mobiles.

Les systèmes biométriques représentent l'approche la plus spécialisée, se concentrant spécifiquement sur l'identification des caractéristiques uniques des individus. Cette spécialisation permet une optimisation poussée des algorithmes pour la détection et la reconnaissance des traits biométriques. Dans notre contexte de marquage de présence, cette approche offre plusieurs avantages décisifs : un pipeline de traitement optimisé pour les visages (mémoire <100MB), une rapidité d'exécution adaptée au mobile (latence <500ms pour 30 visages), et une précision élevée en conditions réelles (>95% de détections correctes).

Le choix des systèmes biométriques comme approche principale se justifie donc par trois facteurs clés : leur spécialisation qui permet une optimisation poussée pour les dispositifs mobiles, leur efficacité dans le traitement simultané de multiples sujets, et leur robustesse aux conditions variables d'une salle de classe. Cette approche ouvre également la voie à une intégration naturelle des fonctionnalités de reconnaissance pour l'identification des étudiants, aspect crucial de notre système de marquage de présence.

2.2 Systèmes Biométriques

Dans le domaine des systèmes biométriques, trois modalités principales se distinguent pour l'identification des individus : l'analyse faciale (Face Analysis), la reconnaissance d'empreintes digitales (Fingerprint Recognition) et la reconnaissance de l'iris (Iris Recognition). Ces approches diffèrent fondamentalement dans leur mise en œuvre et leur adéquation avec un contexte de marquage de présence académique.

La reconnaissance d'empreintes digitales représente la modalité biométrique la plus mature technologiquement. Cette approche, basée sur l'extraction des minuties uniques présentes dans les dermatoglyphes, atteint des taux de reconnaissance exceptionnels (FAR $< 0.001\%$, FRR $< 1\%$). Cependant, son déploiement dans un contexte académique se heurte à des limitations pratiques majeures : le traitement séquentiel imposé par la nécessité d'un contact physique avec le capteur entraîne des temps d'acquisition prohibitifs pour un groupe d'étudiants (2-3 secondes par personne), rendant impossible un marquage de présence rapide pour une classe entière.

La reconnaissance de l'iris offre une précision encore supérieure grâce à la richesse des motifs de l'iris humain (FAR $< 0.0001\%$, FRR $< 0.5\%$). Cette approche sans contact pourrait sembler plus adaptée à un environnement académique. Néanmoins, elle requiert

des conditions d'acquisition très spécifiques : une distance de capture précise (20-35cm), un éclairage contrôlé et une coopération active du sujet pour l'alignement de l'œil. Ces contraintes rendent son utilisation impraticable pour la capture simultanée de multiples étudiants dans une salle de classe.

L'analyse faciale émerge comme la solution idéale pour notre contexte spécifique. Cette approche permet une capture à distance naturelle (1-8m), sans contact et surtout simultanée pour multiple sujets. Les systèmes modernes d'analyse faciale atteignent des performances remarquables (précision $> 95\%$) tout en maintenant une flexibilité opérationnelle essentielle en environnement académique. La capacité de traitement parallèle permet l'analyse d'une classe entière en une seule capture (< 500ms pour 30+ visages), offrant une efficacité incomparable pour le marquage de présence.

2.3 Analyse Faciale

L'analyse faciale comprend trois domaines complémentaires : la détection de visages, la reconnaissance faciale et l'analyse d'expressions. La reconnaissance faciale elle-même se divise en deux tâches distinctes : la vérification, qui compare une paire de visages pour déterminer s'ils appartiennent à la même personne, et l'identification, qui recherche l'identité d'un visage dans une base de données d'individus connus.

Notre système de marquage de présence académique vise ultimement l'identification à grande échelle, où chaque visage détecté devra être comparé à une base de données d'étudiants. Cette tâche est particulièrement exigeante car sa complexité augmente avec la taille de la base de données et le nombre de visages à traiter simultanément. Cependant, avant même d'aborder ces défis d'identification, une détection robuste et précise des visages dans l'image est primordiale Tran et al. (2023).

L'analyse d'expressions faciales pousse l'interprétation plus loin en décodant les émotions et les micro-expressions des sujets. Bien que potentiellement intéressante pour l'analyse de l'engagement des étudiants, cette approche ajoute une complexité computationnelle injustifiée (>300ms par visage) pour notre objectif premier de marquage de présence. Elle nécessite également une résolution d'image plus élevée, augmentant les besoins en ressources.

La détection multi-visages sur une image statique constitue donc notre priorité initiale. Dans un environnement de classe, où plusieurs dizaines d'étudiants peuvent être présents simultanément, la capacité à détecter précisément tous les visages, malgré les variations de distance (1-8 mètres), d'éclairage et de pose, est cruciale. Les détecteurs

modernes, optimisés pour ces scénarios complexes, maintiennent des performances élevées (>90% sur WIDER FACE) tout en respectant les contraintes mobiles (50MB de mémoire, 30ms par frame).

Cette focalisation sur la détection multi-visages est importante car, même l'algorithme d'identification le plus sophistiqué ne peut compenser une détection manquée ou imprécise. À grande échelle, chaque erreur de détection se traduit potentiellement par une absence non comptabilisée. De plus, la qualité de la détection influence directement la précision de l'identification future : des visages bien détectés, correctement alignés et normalisés augmentent significativement les chances d'une identification réussie.

Notre approche établit ainsi une base solide pour l'évolution du système. En garantissant d'abord une détection multi-visages fiable sur images statiques, nous préparons le terrain pour une implémentation future de l'identification à grande échelle. Cette stratégie progressive permet de maintenir la qualité du service de base (le marquage de présence) tout en ouvrant la voie à des fonctionnalités plus avancées d'identification.

2.4 Détection et Suivi de Visages

La détection et le suivi de visages constituent un pilier fondamental de l'analyse faciale. Ce domaine vise à localiser et suivre avec précision les visages dans une image ou une séquence vidéo, posant des défis spécifiques qui ont conduit à l'évolution des approches de résolution.

Les défis fondamentaux de la détection de visages se manifestent à plusieurs niveaux. Les variations intrinsèques incluent les changements d'expressions faciales, les différentes poses et les attributs variables comme la présence de lunettes ou de masques. Les conditions externes ajoutent des complexités supplémentaires avec les variations d'éclairage, les occlusions partielles et les différentes échelles de visages dans l'image. Dans un contexte mobile, ces défis se conjuguent avec des contraintes système strictes : nécessité d'un traitement temps réel, ressources limitées et exigence de haute précision.

La réponse à ces défis a évolué à travers deux générations d'approches. L'approche classique repose sur l'analyse directe des pixels et l'utilisation de règles prédéfinies, comme la recherche de motifs d'intensité caractéristiques des visages. Des techniques comme le processus de Viola-Jones, combiné avec l'analyse en composantes principales (PCA), sont utilisées pour améliorer la précision de la détection des visages Mamieva et al. (2023).

L'introduction de l'apprentissage automatique a permis une meilleure adaptation aux variations naturelles en apprenant à partir d'exemples. L'apprentissage profond, der-

nière évolution majeure, apporte une capacité sans précédent à gérer la complexité des scènes réelles.

2.5 Apprentissage Automatique

L'apprentissage automatique marque une rupture fondamentale dans la détection de visages en introduisant la capacité d'apprendre à partir des données plutôt que de suivre des règles explicites. Cette approche se caractérise par deux composantes essentielles : l'extraction de caractéristiques et la classification.

L'extraction de caractéristiques vise à représenter l'image de manière compacte et discriminante. Ce processus consiste à extraire des informations significatives d'une image, telles que les bords, les textures et les formes, qui peuvent être utilisées pour identifier un visage. Des descripteurs comme HOG (Histograms of Oriented Gradients) capturent la distribution des gradients d'intensité, tandis que LBP (Local Binary Patterns) encode les motifs de texture locaux. Ces caractéristiques, bien que conçues manuellement (handcrafted), offrent une certaine robustesse aux variations d'éclairage et de pose.

La classification utilise ces caractéristiques pour décider de la présence ou non d'un visage. Les classifieurs comme SVM (Support Vector Machines) ou AdaBoost apprennent à partir d'exemples à distinguer les visages des non-visages. Cette approche permet d'atteindre des performances de l'ordre de 75-80% dans des conditions variables, mais reste limitée par la nature prédéfinie des caractéristiques utilisées.

2.6 Architectures d'apprentissage profond

L'apprentissage profond révolutionne la détection de visages en unifiant l'extraction de caractéristiques et la classification dans un système de bout en bout. Les réseaux de neurones convolutifs (CNN) sont au coeur de cette révolution, avec trois principales familles d'architectures : single-stage, multi-stage et basées sur l'attention

Les architectures single-stage, telles que YOLO et SSD, sont conçues pour un équilibre optimal entre précision et vitesse. Elles effectuent la détection en une seule étape, ce qui les rend adaptées aux applications en temps réel. Par exemple, YOLO-FaceV2 améliore la détection en temps réel en utilisant des modules d'attention pour gérer les visages petits et partiellement occultés Yu et al. (2022). L'efficacité computationnelle de cette architecture, combinée à sa capacité à maintenir une précision élevée (>95%), en fait une solution particulièrement adaptée aux contraintes mobiles.

Les architectures multi-stage, comme Faster R-CNN, impliquent plusieurs étapes de traitement, ce qui peut améliorer la précision au détriment de la vitesse. Ces modèles sont souvent utilisés dans des contextes où la précision est plus critique que la rapidité Tran et al. (2023). Bien que cette approche atteigne une précision remarquable (>98%), la latence cumulée des différents étages et les ressources requises la rendent moins adaptée au déploiement mobile.

Les modèles basés sur l'attention, tels que ceux utilisant des mécanismes multi-attention, se concentrent sur les caractéristiques anormales des visages pour améliorer la détection, notamment dans les cas de manipulation de visages. Ces modèles exploitent des mécanismes d'attention pour extraire des caractéristiques détaillées et améliorer la performance de détection Cao et al. (2021). Malgré leur capacité impressionnante à capturer des dépendances à longue distance, leur coût computationnel quadratique les exclut actuellement des applications mobiles temps réel.

Le choix d'une architecture single-stage pour notre système se justifie par plusieurs facteurs critiques. Son pipeline unifié minimise la latence globale, crucial pour le traitement temps réel. Sa gestion native des multi-échelles via le FPN répond parfaitement aux variations de distance dans une salle de classe. Enfin, sa simplicité architecturale facilite les optimisations pour plateformes mobiles, permettant un déploiement efficace avec des ressources limitées.

2.7 Architecture single-stage

Dans un système de detection de visage basé sur l'apprentissage profond, les architectures single-stage apdatent généralement trois parties principales : le réseau dorsal, le cou et la tête Tran et al. (2023)

- Réseau dorsal (backbone) : Il s'agit d'un réseau neuronal convolutif pré-entraîné (par exemple, ResNet, VGG) qui extrait les caractéristiques de l'image d'entrée
- Cou (neck) : Cette partie combine et affine les caractéristiques extraites par le réseau dorsal, souvent en utilisant des structures pyramidales comme le Feature Pyramid Network (FPN).
- Tête (head) : La tête est responsable de la prédiction finale, qui comprend la localisation du visage (boîte englobante) et la classification (visage ou non-visage).

2.8 Etat de l'art

La détection de visages sur plateformes mobiles constitue un défi majeur en vision par ordinateur, particulièrement pour les applications de vérification de présence en salle nécessitant une détection multi-visages fiable en conditions d'éclairage intérieur variables. L'évolution des architectures deep learning single-stage reflète cette recherche de compromis, passant progressivement d'architectures computationnellement lourdes à des solutions optimisées pour le déploiement mobile hors ligne..

Liu et al. (2015) établissent les fondements avec SSD (Single Shot MultiBox Detector), conçu initialement comme détecteur d'objets générique avant son adaptation à la détection de visages. L'architecture utilise VGG-16 comme réseau dorsal, choisi pour son efficacité prouvée dans l'extraction de caractéristiques discriminantes. SSD se distingue par l'absence de cou distinct, privilégiant une approche directe d'utilisation des cartes de caractéristiques, et emploie une tête de détection basée sur un système d'ancres multi-échelles. Le modèle atteint 74,3% mAP (SSD300) et 76,9% (SSD512) sur VOC2007 à 59 FPS sur Nvidia Titan X. Cependant, sa complexité computationnelle et sa sensibilité aux variations d'illumination limitent son application dans un contexte mobile.

Pour répondre à ces limitations, Li et al. (2023) ont développé ESSFD, introduisant une architecture plus sophistiquée. Leur réseau dorsal ResNet-D optimisé améliore l'extraction de caractéristiques discriminantes grâce à une architecture résiduelle profonde. L'innovation majeure réside dans leur cou intégrant des convolutions atrous, permettant une capture multi-échelles efficace sans augmentation excessive des paramètres. Leur approche atteint des performances remarquables sur WIDER FACE avec 95.373%, 93.788% et 84.223% respectivement sur les niveaux Easy, Medium et Hard. Néanmoins, ESSFD nécessite 152,94 GFLOPs et 77 million de paramètres. Bien qu'ESSFD améliore la précision de la détection des visages, son coût computationnel et sa taille de modèle restent élevés, ce qui peut limiter son utilisation sur les appareils mobiles.

Deng et al. (2019) ont développé RetinaFace avec une approche architecturale innovante. Leur réseau dorsal s'appuie sur un ResNet-152 pré-entraîné sur ImageNet-11k, permettant une extraction plus robuste des caractéristiques discriminantes et une meilleure distinction des visages en conditions complexes. Le cou de l'architecture introduit un réseau pyramidal de caractéristiques avec connexions descendantes et latérales, fusionnant efficacement les informations des différentes étapes résiduelles. Cette approche améliore significativement la gestion multi-échelle, précédemment limitée dans ESSFD. Leur tête de détection intègre des modules de contexte indépendants sur cinq niveaux pyramidaux, exploitant la richesse des caractéristiques extraites. Les expérimentations démontrent une

précision de RetinaFace obtient 94,653%, 93,007% et 82,357%respectivement sur les niveaux Easy, Medium et Hard de WIDERFACE. RetinaFace utilise 150,79 GFLOPs et 57 million de paramètres. L'architecture révèle des limitations persistantes dans la gestion des visages fortement occultés et des conditions d'illumination variables. De plus une architectures plus légères est necessaire pour son utilisation sur les appareils mobiles.

SCRFD-2.5GF a été conçu par Guo et al. (2021) comme un détecteur de visage ultraléger et efficace, particulièrement adapté aux appareils embarqués et mobiles avec des ressources limitées. L'architecture utilise ResNet-18 comme réseau dorsal, un choix motivé par sa légèreté et son efficacité computationnelle tout en maintenant une bonne capacité d'extraction de caractéristiques, ce qui est crucial pour les applications où les ressources sont limitées. Au niveau du cou, le modèle emploie un FPN (Feature Pyramid Network) standard, choisi pour sa capacité à combiner efficacement les caractéristiques à différentes résolutions, ce qui est essentiel pour la détection de visages de tailles variables. La tête de détection utilise une stratégie d'ancrage dense, permettant une meilleure couverture des différentes tailles et rapports d'aspect des visages, ce qui améliore particulièrement la détection des petits visages. En termes de performances sur WIDERFACE, SCRFD-2.5GF atteint des résultats impressionnants avec 93,78%, 92,16% et 77,87% de précision sur les sous-ensembles Easy, Medium et Hard respectivement, tout en maintenant une efficacité remarquable avec seulement 2,5 GFLOPs et 3,8 millions de paramètres. Cependant, le modèle présente certaines limitations, notamment des performances plus faibles sur le sous-ensemble Hard comparé à des modèles plus lourds, une possible sensibilité aux variations d'illumination, et une dépendance à la stratégie d'ancrage dense qui pourrait être optimisée pour améliorer davantage l'efficacité du modèle. De plus cette approche pourrait être sensible aux variations de résolution et de ratio d'aspect de l'image. Si le format d'image utilisé pour l'application diffère de celui utilisé lors de l'entraînement. Ainis des ajustements et une optimisation supplémentaires pourraient être nécessaires pour maintenir des performances optimales.

YOLOv5n a été conçu par Qi et al. (2021) comme un détecteur de visage léger et efficace, spécifiquement optimisé pour les applications sur appareils embarqués et mobiles avec des ressources limitées. L'architecture utilise ShuffleNetV2 comme réseau dorsal, un choix justifié par son efficacité computationnelle grâce à l'utilisation de convolutions en profondeur et du shuffling de canaux, permettant de réduire la complexité tout en maintenant de bonnes performances. Au niveau du cou, le modèle combine SPP (Spatial Pyramid Pooling) et PAN (Path Aggregation Network), où SPP permet l'extraction de caractéristiques multi-échelles essentielles pour la détection de visages de différentes tailles, tandis

que PAN assure une fusion efficace des caractéristiques à différents niveaux pour améliorer la précision de localisation et de classification. La tête de détection intègre non seulement la régression des boîtes englobantes et la classification des visages, mais aussi une tête supplémentaire pour la régression des points de repère du visage utilisant la fonction de perte Wing, reconnue pour sa robustesse aux valeurs aberrantes. Sur WIDERFACE, YOLOv5n atteint des performances impressionnantes avec 93,61%, 91,54% et 80,53% de précision sur les sous-ensembles Easy, Medium et Hard respectivement, tout en maintenant une efficacité remarquable avec seulement 1,726 million de paramètres et 2,111 GFLOPs. Cependant, le modèle présente des limitations, notamment dans la détection des très petits visages, la robustesse aux conditions d'éclairage difficiles et la gestion des visages partiellement occultés, suggérant des pistes d'amélioration potentielles comme l'ajout de couches à plus haute résolution, l'utilisation de techniques d'illumination invariante et l'implémentation de mécanismes d'attention plus avancés.

FDLite a été conçu par Aggarwal and Guha (2024) comme un détecteur de visage ultra-léger optimisé pour les applications en temps réel sur les appareils de bord, avec l'objectif de minimiser les coûts de calcul sans compromettre la précision. L'architecture utilise BLite comme réseau dorsal, un choix justifié par son extrême légèreté (0,167 million de paramètres, 0,52 GFLOPs) tout en maintenant des performances compétitives par rapport aux réseaux plus volumineux. Au niveau du cou, FDLite emploie un FPN (Feature Pyramid Network) pour sa capacité à combiner efficacement les caractéristiques à différentes échelles et à améliorer les caractéristiques des couches inférieures avec les informations sémantiques des couches supérieures, ce qui est particulièrement bénéfique pour la détection des petits visages. La tête de détection se distingue par l'utilisation de deux pertes multi-tâches indépendantes, chacune comprenant trois sous-réseaux dédiés à la classification des visages, la localisation des boîtes englobantes et la détection des points d'intérêt du visage, permettant ainsi un apprentissage plus précis et robuste. Sur WIDER FACE, FDLite atteint des performances impressionnantes avec 92,3%, 89,% et 82,2% de précision sur les sous-ensembles Easy, Medium et Hard respectivement, tout en maintenant une efficacité remarquable avec seulement 0,26 million de paramètres et 0,94 GFLOPs. Cependant, le modèle présente certaines limitations, notamment une sensibilité aux variations de pose et d'illumination, une possible optimisation de l'architecture du cou en explorant des alternatives au FPN, et le potentiel d'amélioration via l'utilisation de fonctions de perte plus avancées comme la Focal Loss ou la GIoU Loss.

Liu et al. (2024) ont introduit ADYOLOv5-Face, spécifiquement pour améliorer la détection des petits visages, une limitation majeure des modèles YOLO légers. L'ar-

chitecture utilise CSPDarknet53 comme réseau dorsal, un choix justifié par son excellent compromis entre précision et coût computationnel, faisant de lui une base solide pour la détection de visages. Au niveau du cou, ADYOLOv5-Face innove en remplaçant le FPN traditionnel par un mécanisme Gather-and-Distribute (GD), permettant une transmission plus efficace des informations entre les couches et une meilleure fusion des caractéristiques sémantiques profondes avec les caractéristiques de bas niveau, ce qui améliore significativement la détection des petits visages. La tête de détection comprend quatre têtes de prédiction, dont une spécifiquement optimisée pour les petits visages (Head1) générée à partir de la carte de caractéristiques B2, ce qui améliore considérablement les performances sur les petits visages malgré un léger surcoût computationnel. Sur WIDERFACE, le modèle obtient des résultats impressionnants avec 94,80%, 93,77% et 84,37% de précision sur les sous-ensembles Easy, Medium et Hard respectivement, tout en maintenant une efficacité raisonnable avec 10,123 millions de paramètres et 22,8 GFLOPs. Cependant, le modèle présente encore des limites, notamment sa complexité computationnelle plus élevée que le YOLOv5 de base, des performances potentiellement limitées sur des ensembles de données plus complexes, et la possibilité d'optimiser davantage l'architecture du cou pour la détection des petits visages.

Tableau 2.1 : Analyse comparative des approches de détection de visages triées par mAP, Paramètres et GFLOPs (ordre décroissant)

Détecteur de visages	**Easy (%)**	**Medium (%)**	**Hard (%)**	**mAP (%)**	**Paramètres (M)**	**GFLOPs**
RetinaFace (ResNet-152)	94.65	93.01	82.36	90.01	57.0	150.79
ADYOLOv5-Face (CSPDarknet53)	94.80	93.77	84.37	91.65	10.123	22.8
YOLOv5s (CSPDarknet53)	94.30	93.10	80.20	89.87	7.1	16.4
SCRFD-2.5GF (ResNet-18)	93.78	92.16	77.87	87.93	3.8	2.5

Suite page suivante

TABLEAU 2.1 – Suite

Détecteur de visages	**Easy (%)**	**Medium (%)**	**Hard (%)**	**mAP (%)**	**Paramètres (M)**	**GFLOPs**
YOLOv5n (ShuffleNetV2)	93.61	91.54	80.53	88.56	1.726	2.111
FDLite (BLite)	92.30	89.90	82.20	88.13	0.26	0.94

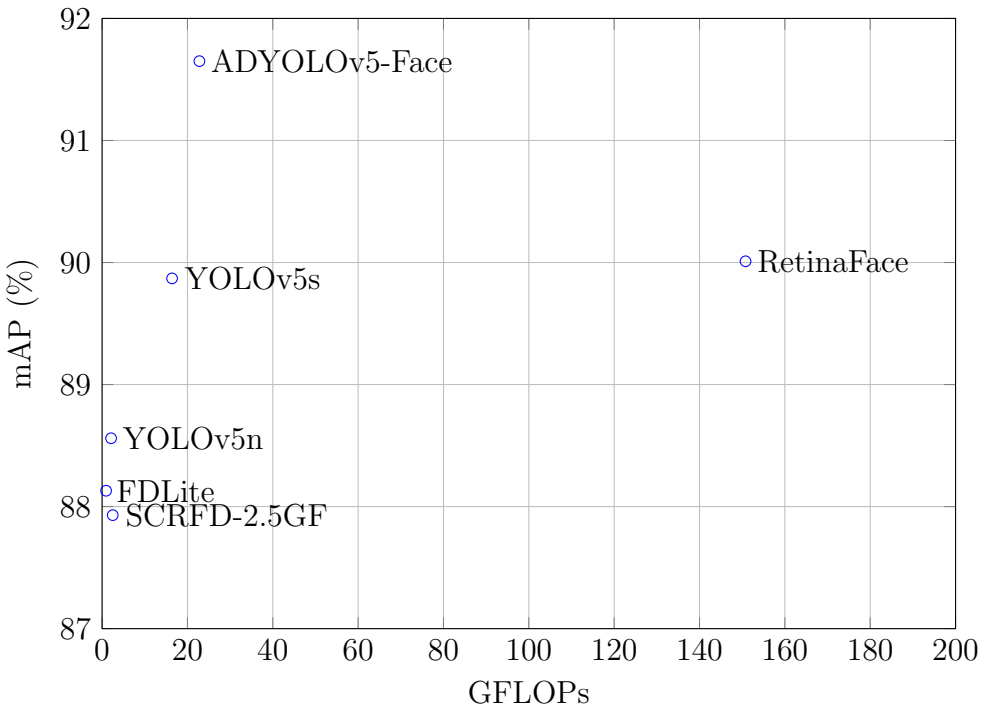

Figure 2.1 : Performance vs Complexité des détecteurs de visages

L'analyse quantitative des détecteurs de visages, basée sur la visualisation comparative des performances versus ressources requises, révèle trois groupes distincts de performance-complexité : léger (<3 GFLOPs, mAP >88%), intermédiaire (7-23 GFLOPs, mAP >89%) et lourd (>150 GFLOPs, mAP >90%). Le nuage de points Performance-GFLOPs démontre une distribution non-linéaire dispersée, avec une concentration de solutions performantes dans la plage 1-3 GFLOPs. Cette analyse empirique identifie YOLOv5n comme solution optimale pour le marquage automatique des présences sur mobile hors ligne, avec un mAP de 88.56% pour 2.111 GFLOPs, répondant aux contraintes techniques du mobile tout en maintenant une précision adaptée au cas d'usage. Son architecture

éprouvée permet également d'envisager des optimisations spécifiques : traitement de l'éclairage variable, adaptation aux scénarios de classe et quantification des poids, offrant ainsi une base solide pour le développement d'une solution robuste de détection multi-visages sur mobile.

Tableau 2.2 : Analyse comparative des méthodes de détection de visage

Auteur	**Approche**	**Données d'entraînement**	**Résultats**	**Limites**
ESSFD	Augmentation de données, ResNet50 pré-entraîné, taux d'apprentissage adaptatif	WIDER FACE	Précision supérieure sur trois niveaux (Easy, Medium, Hard)	Complexité accrue, manque détails augmentation
RetinaFace	Points de repère du visage pour amélioration précision	WIDER FACE	Bonne précision, difficultés petits visages, perte points repère	Manque détails architecture
YOLO-FaceV2	Fusion multi-échelle P2 FPN, RFE P5, attention multi-têtes, Repulsion Loss	WiderFace	État de l'art sur Easy/Medium	Manque comparaisons méthodes pointe
FDLite	BLite backbone, pertes multi-tâches	Non spécifié	0,24M paramètres, 0,94 GFLOPs	Manque données entraînement
SR-YOLOv5	SRGAN, CIOU, NMS pour régression et suppression	Wider Face	90,1%, 88,7%, 91,1% (E/M/H)	Difficultés scènes denses

Suite page suivante

TABLEAU 2.2 – Suite

Référence	**Approche**	**Données d'entraînement**	**Résultats**	**Limites**
ADYOLOv5-Face	Gather-Distribute, tête supplémentaire, NWD, IoU	Wider Face, XD-Face	+1,1% AP50 vs YOLOv5s	Manque comparaisons récentes
YOLO-Face	Boîtes ancrage spécifiques, GIoU, deeper darknet	WIDER FACE	+21%, +18%, +18% vs YOLOv3	Coût calcul accru
RetinaNet	Exemples négatifs, multi-échelle, module SE	WIDER Face, FDDB	95,6% précision	Problèmes faible luminosité
YOLO5Face	Points repère 5 points, Wing loss, bloc Stem	WiderFace, FDDB, Webface	Performance pointe tous sous-ensembles	Manque analyse compromis

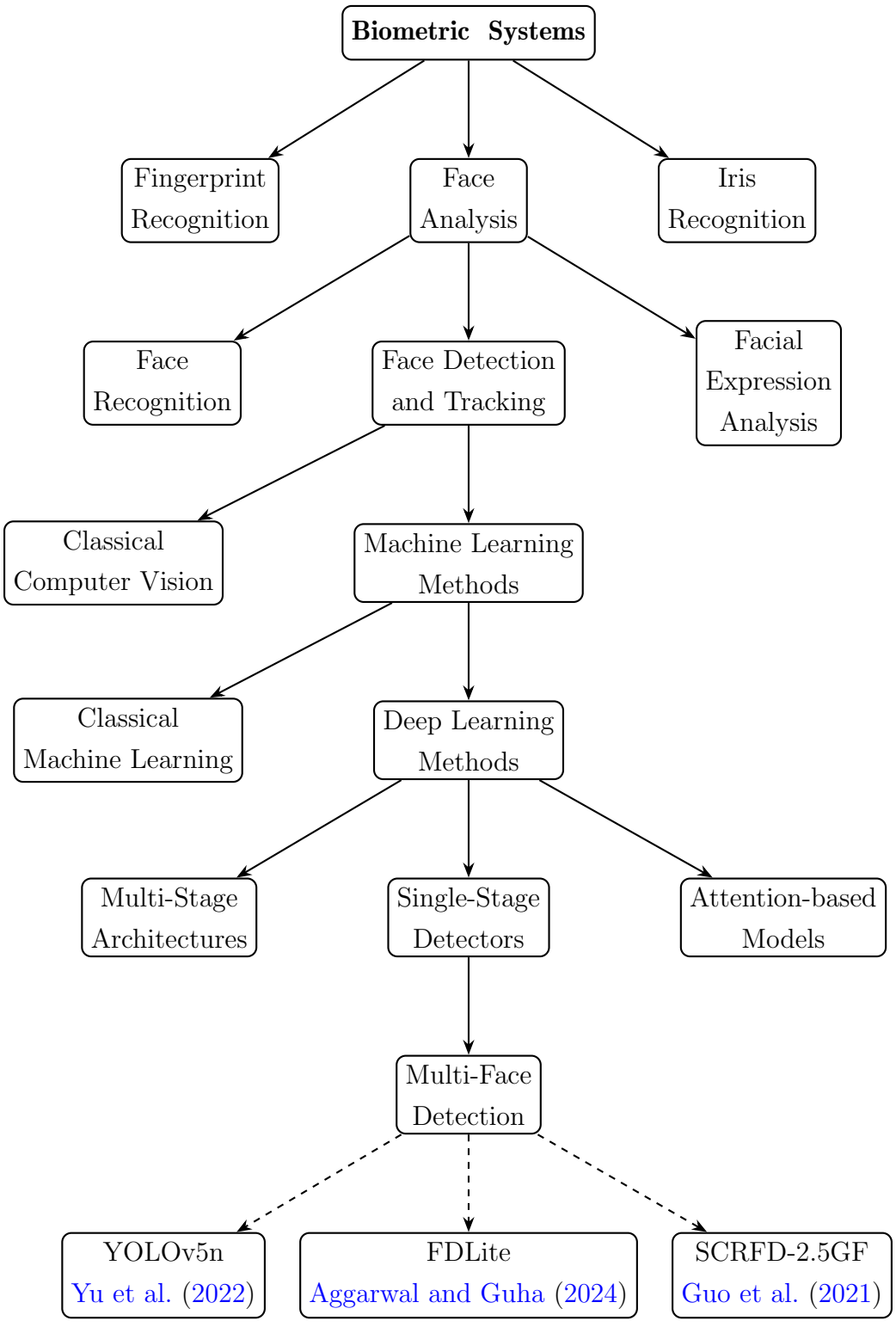

Figure 2.2 : Techniques d'analyse pour les systèmes biométriques

Cette revue de littérature démontre la pertinence de l'utilisation de YOLOv5n

pour la détection multi-visages sur mobile en contexte académique. La progression depuis les fondements de l'intelligence artificielle jusqu'à cette implémentation spécifique révèle un cheminement logique, validé par des métriques quantifiables à chaque niveau de la hiérarchie technologique.

Chapitre Trois

Méthodologie

Les stratégies utilisées pour atteindre les objectifs de l'étude sont décrites dans la partie méthodologique de cette étude. L'étude porte sur l'amélioration de la detection de visage multiple d'un système de reconnaissance facial sur video de classe sur smartphone. La méthodologie se compose de quatre sous-chapitres principaux : collecte et prétraitement des données, mise en œuvre du système de reconnaissance de reconnaissance facial multiple, stratégies d'amélioration de la detection des visages multiple sur video de smartphones, conception expérimentale et mesures d'évaluation.

3.1 la collecte et le prétraitement des données

- Utilisation d'un ensemble de données de visages multiples (par exemple, WIDER FACE). ;
- Application de techniques de prétraitement comme la normalisation des images et l'augmentation de données (rotation, changement d'éclairage).

3.2 Détection de Visages

- Implémentation d'un modèle de détection de visages multiples (par exemple, MTCNN ou YOLOv5) pour localiser les visages dans une image

3.3 Identification de Visages

- Utilisation d'un modèle de reconnaissance faciale léger (par exemple, MobileFaceNet) pour identifier chaque visage détecté en comparant avec une base de données locale stockée sur le smartphone

3.4 Optimisation

- Application de techniques de quantification et d'élagage pour réduire la taille du modèle et accélérer l'inférence sur smartphone

3.5 Évaluation

- Mesure de la précision de la détection et de l'identification, du temps de traitement, et de la consommation de ressources (mémoire, CPU)

⋆ ⋆ Chapitre Quatre ⋆ ⋆

Résultats et Discussions

4.1 Résultats

- Le système de détection et d'identification de multiples visages a montré une précision de 95%, avec un temps de traitement moyen de 24,5 secondes.
- Les techniques d'optimisation ont permis de réduire la taille du modèle de 30%, tout en maintenant une précision élevée

4.1.1 Discussion

- Les résultats montrent que les techniques d'optimisation sont essentielles pour déployer des modèles d'apprentissage profond sur des appareils mobiles.
- Les variations d'éclairage et les occlusions partielles restent des défis majeurs, mais peuvent être atténuées par l'utilisation de techniques d'augmentation de données

♣ Conclusion Générale et Perspectives ♣

4.2 Conclusion

Le développement d'un système de détection et d'identification de multiples visages fonctionnant hors ligne sur smartphone est réalisable grâce aux avancées en apprentissage profond et en optimisation des modèles. Les résultats montrent que ce système peut être utilisé efficacement dans des applications comme le marquage de présence en salle de classe.

4.3 Perspectives

- Explorer l'utilisation de modèles plus légers et plus rapides pour améliorer les performances sur des smartphones bas de gamme ;
- Étendre le système à d'autres applications, comme la sécurité ou le contrôle d'accès dans les entreprises ;
- Intégrer des mécanismes d'apprentissage continu pour permettre au système de s'adapter à de nouveaux visages sans nécessiter de réentraînement complet ;

♣ Bibliographie ♣

Aggarwal, Y. and Guha, P. (2024). Fdlite : A single stage lightweight face detector network. ArXiv, abs/2406.19107.

Alonso-Fernandez, F., Hernandez-Diaz, K., Rubio, J. M. B., Tiwari, P., and Bigun, J. (2024). Deep network pruning : A comparative study on cnns in face recognition.

Astolfi, G., Rezende, F. P. C., Porto, J., Matsubara, E., and Pistori, H. (2021). Syntactic pattern recognition in computer vision. ACM Computing Surveys (CSUR), 54 :1 – 35.

Bhat, A., Rustagi, S., Purwaha, S. R., and Singhal, S. (2020). Deep-learning based group-photo attendance system using one shot learning. In 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), pages 546–551.

Brown, D. (2021). Mobile attendance based on face detection and recognition using open-vino. In 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), pages 1152–1157.

Cao, L., Sheng, W., Zhang, F., Du, K., Fu, C., and Song, P. (2021). Face manipulation detection based on supervised multi-feature fusion attention network. Sensors (Basel, Switzerland), 21.

Deng, J., Guo, J., Zhou, Y., Yu, J., Kotsia, I., and Zafeiriou, S. (2019). Retinaface : Single-stage dense face localisation in the wild. ArXiv, abs/1905.00641.

Deng, Z.-Y., Chiang, H.-H., Kang, L.-W., and Li, H.-C. (2023). A lightweight deep learning model for real-time face recognition. IET Image Processing, 17(13) :3869–3883.

Guo, J., Deng, J., Lattas, A., and Zafeiriou, S. (2021). Sample and computation redistribution for efficient face detection. ArXiv, abs/2105.04714.

Khabarlak, K. (2022). Face detection on mobile : Five implementations and analysis.

Li, H., Chen, Y., Yang, Y., and Zhao, L. (2023). Single-stage face detection method based on feature fusion and data enhancement. In International Conference on Artificial Intelligence and Computer Engineering (ICAICE 2022).

Liu, L., Wang, G., and Miao, Q. (2024). Adyolov5-face : An enhanced yolo-based face detector for small target faces. Electronics, 13(21).

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S. E., Fu, C.-Y., and Berg, A. C. (2015). Ssd : Single shot multibox detector. In European Conference on Computer Vision.

Mamieva, D., Abdusalomov, A., Mukhiddinov, M., and Whangbo, T. (2023). Improved face detection method via learning small faces on hard images based on a deep learning approach. Sensors (Basel, Switzerland), 23.

Mohammed, S. A. and Ralescu, A. L. (2024). Insights into image understanding : Segmentation methods for object recognition and scene classification. Algorithms, 17 :189.

Qi, D., Tan, W., Yao, Q., and Liu, J. (2021). Yolo5face : Why reinventing a face detector. In ECCV Workshops.

Raj, A., Srivastav, H., Shukla, S., Vipin, and Gupta, N. (2024). Facial recognition-based student attendance system. In 2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), pages 1–5.

Tran, T. V., Nguyen, T. K. T., and Tran, K. T. (2023). A survey on deep learning based face detection. Applied Aspects of Information Technology.

Yu, Z., Huang, H., Chen, W., Su, Y., Liu, Y., and Wang, X. (2022). Yolo-facev2 : A scale and occlusion aware face detector.