République Algérienne Démocratique et Populaire
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique
Université Abbes Laghrour Khenchela
Faculté des Sciences et de la Technologie
Département des Mathématiques et Informatique

N° d'order
Serie :

Thèse pour l'obtention de diplôme
Doctorat (LMD)
Spécialité : Informatique
Laboratoire de recherche : Ingénerie des COnnaissances et Sécurité Informatique (ICOSI)

Une approche deep learning pour la reconstruction de visage 3D

Présenté par

Malah Mehdi

Soutenue le 15/01/2024

Membres de Jury :

Siam Abderrahim	MCA. - Université de Khenchela	Président
Hemam Mounir	Prof. - Université de Khenchela	Directeur de thèse
Abbas Fayçal	MCA - Université de Khenchela	Co-Directeur de thèse
Babahenini Mohamed Chaouki	Prof. - Université de Biskra	Examinateur
Abdelhadi Adel	MCA - Université de Batna 2	Examinateur
Bardou Dalal	MCA - Université de Khenchela	Examinatrice
Djezzar Meriem	MCA - Université de Khenchela	Invitée

Dédicace

Ma famille et mes amis

Remerciements

En premier lieu, j'aimerais remercier le Tout-Puissant d'avoir fait preuve de bienveillance en me donnant le courage et la volonté nécessaires pour mener à bien cette thèse.

Je tiens tout d'abord à exprimer ma profonde gratitude envers mon directeur et co-directeur de thèse,le Pr. Mounir Hemam et le Dr. Fayçal ABBAS et . J'ai pu bénéficier de leur grande expertise en recherche ainsi que de la qualité de leur encadrement tout au long de cette thèse. Leurs disponibilités et conseils ont été d'une grande valeur.

Je désire également exprimer toute ma reconnaissance envers mes chers parents qui m'ont soutenu et encouragé tout au long de ma vie et de mes études.

Mes remerciements les plus sincères vont à tous mes amis pour leur disponibilité, leurs conseils précieux et leurs remarques constructives qui nous ont permis d'améliorer la qualité de ce travail.

Je tiens également à exprimer toute ma gratitude envers les membres du jury d'avoir accepté d'évaluer ce travail.

Résumé

La modélisation précise des visages est essentielle dans les industries du divertissement, de la médecine, de la sécurité et de l'interaction homme-machine. Les techniques traditionnelles de reconstruction 3D des visages sont confrontées à des défis majeurs, tels que les variations d'expression, de pose, d'illumination, et d'occlusion. Ces défis ont conduit à l'exploration d'approches innovantes basées sur le Deep Learning. Cette thèse présente une approche révolutionnaire qui exploite les réseaux de neurones génératifs (GAN) pour résoudre le défi complexe de la reconstruction 3D des visages à partir d'une seule image 2D.

Nous proposons une nouvelle approche pour la reconstruction de la géométrie du visage, tout en incluant l'expression faciale et la position. En utilisant une seule image 2D et les point de repère correspondant en entrée du générateur, nous parvenons à générer une géométrie du visage avec une seule branche de modèle, ce qui simplifie considérablement le processus tout en préservant la qualité des résultats. Nous avons évalué rigoureusement notre approche de reconstruction 3D des visages en effectuant une comparaison exhaustive avec les méthodes de l'état de l'art. Cette évaluation s'est appuyée à la fois sur des mesures quantitatives et qualitatives. Nos résultats démontrent de manière convaincante que notre modèle génère des meshes de visage d'une grande précision, surpassant les méthodes existantes.

Mots-clés : Reconstruction 3D à partir d'une seule image - Reconstruction faciale - Réseaux antagonistes génératifs - Réseaux de convolution graphique.

Abstract

Precise modeling of faces is crucial in the entertainment, medical, security, and human-machine interaction industries. Traditional 3D face reconstruction techniques face major challenges, such as variations in expression, pose, illumination, and occlusion. These challenges have led to the exploration of innovative approaches based on Deep Learning. This thesis presents a revolutionary approach that leverages Generative Adversarial Networks (GANs) to tackle the complex challenge of 3D face reconstruction from a single 2D image.

We propose a novel approach for reconstructing facial geometry while incorporating facial expression and position. By using a single 2D image and the corresponding landmarks as input to the generator, we manage to generate facial geometry with a single model branch, significantly simplifying the process while preserving the quality of the results. We rigorously evaluated our 3D face reconstruction approach by conducting a comprehensive comparison with state-of-the-art methods. This evaluation relied on both quantitative and qualitative measures. Our results compellingly demonstrate that our model generates highly accurate face meshes, surpassing existing methods.

Keywords: Single image 3D reconstruction - Face reconstruction - Generative adversarial networks - Graph convolution networks.

ملخص

النمذجة الدقيقة للوجوه مهمة في الصناعات الترفيهية والطب والأمن كذلك كأهميتها في التفاعلات ما بين الإنسان والآلة. تواجه تقنيات إعادة إنشاء الوجه ثلاثي الأبعاد التقليدية تحديات كبيرة، مثل تغيرات تعابير ووضعيات الوجه وكذلك الإضاءة والتعتيم. هذه التحديات أدت إلى استكشاف نهج مبتكرة استنادًا على التعلم العميق. تقدم هذه الأطروحة نهجًا جديدًا يستخدم شبكة الأعصاب التوليدية (GANs) لمواجهة التحديات المعقدة لإعادة بناء الوجه ثلاثي الأبعاد انطلاقا من صورة واحدة ثنائية الأبعاد.

نقترح نهجًا جديدًا لإعادة بناء هندسة الوجه مع مراعاة تعابير ووضعيات الوجه. انطلاقا من صورة ثنائية الأبعاد للوجه والنقط الاستدلالية الخاصة بها كإدخال للمولد، نجحنا في توليد نموذج ثلاثي الأبعاد للوجه باستخدام شبكة توليدية واحدة، مما يبسط بشكل كبير العملية الحسابية وفي الوقت نفسه يحافظ على جودة النتائج.

قمنا بتقييم نهجنا لإعادة بناء الوجه ثلاثي الأبعاد بشكل دقيق من خلال إجراء مقارنة شاملة مع الأساليب الحديثة. تظهر نتائجنا بشكل مقنع أن نموذجنا يُنتِج شبكات وجه دقيقة للغاية، متفوقة على الأساليب والطرق الحديثة.

الكلمات المفتاحية:
إعادة البناء ثلاثية الأبعاد من صورة واحدة، إعادة بناء الوجه، شبكات الأعصاب التوليدية، الشبكات التلافيفية للرسم البياني.

Table des Matières

Liste des Figures

Liste des Tableaux

Liste des Abbréviations

3DMM – 3D Morphable Model.

CD – Chamfer Distance.
CGAN – Conditional Generative Adversarial Network.
CNN – Convolutional Neural Network.

DCGAN – Deep Convolutional Generative Adversarial Network.
DECA – Detailed Expression Capture and Animation.

EMD – Earth Mover's Distance.

GAN – Generative Adversarial Network.
GCN – Graph Convolutional Network.

LS3DMM – Large Scale 3D Morphable Models.

PRNet – Joint 3D Face Reconstruction and Dense Alignment with Position Map Regression.

ReLU – Rectified Linear Unit.
ResNet – Residual Network.

SFM – Structure From Motion.
SVD – Singular Value Decomposition.

VGG – Visual Geometry Group.

WGAN – Wasserstein Generative Adversarial Network.

1

Introduction Générale

1.1 Contexte de la Recherche

La reconstruction tridimensionnelle des visages constitue un domaine de recherche fondamental et une application pratique de plus en plus cruciale, notamment dans les domaines des jeux vidéo, du cinéma d'animation et de l'industrie esthétique. L'objectif principal de cette thèse est de présenter une nouvelle méthode de reconstruction 3D à l'aide de modèles génératifs basés sur des réseaux de neurones antagonistes.

Cette approche a le potentiel de révolutionner la manière dont nous modélisons la géométrie du visage 3D en utilisant une seule image bidimensionnelle et ses points de repère associés. Traditionnellement, la reconstruction tridimensionnelle des visages a été un défi complexe en raison de diverses limitations inhérentes aux techniques classiques. Ces limites incluent la sensibilité à la variation de l'illumination, de la position, de l'occultation et de l'expression faciale. Les méthodes traditionnelles se sont heurtées à des difficultés majeures lorsqu'il s'agissait de capturer avec précision la géométrie du visage dans des conditions de prise de vue non contrôlées.

Cependant, avec l'avènement de l'apprentissage profond et l'essor des processeurs graphiques, de nouvelles possibilités se sont ouvertes. Notre thèse repose sur l'hypothèse que les réseaux de neurones génératifs, en particulier les modèles Generative Adversarial Network (GAN) [1], peuvent offrir une solution innovante à ces problèmes. Notre recherche vise à utiliser ces modèles pour créer des représentations tridimensionnelles détaillées des visages à partir d'une unique image en deux dimensions et des points de repère associés.

Le cœur de notre approche repose sur l'architecture de notre générateur et de notre discriminateur, tous deux basés sur des couches de convolution graphique [2]. Notre approche consiste à utiliser une image 2D et ses points de repère du visage en tant qu'entrée du générateur, ce qui permet de capturer à la fois la géométrie tridimensionnelle du visage ainsi que son expression. En d'autres termes, nous cherchons à créer un modèle unique capable de générer une représentation 3D du visage qui prend en compte à la fois la position et l'expression faciale. Les résultats préliminaires de notre recherche montrent que notre modèle génère des reconstructions de visage de haute qualité, notamment en termes d'expression faciale, par rapport aux objets de référence.

Notre contribution peut être résumée comme suit :

- Nous présentons une méthode novatrice pour la reconstitution de la géométrie tridimensionnelle du visage, en prenant en compte l'expression et la position, à partir d'une unique image en deux dimensions et ses points de repère du visage, en utilisant les réseaux de neurones génératifs.

- Nous introduisons une architecture de discriminateur basée sur les GCN pour améliorer la précision de la représentation 3D du visage générée.

- Nous exploitons les points de repère en tant qu'entrée du générateur pour produire une représentation 3D du visage prenant en compte à la fois la géométrie et l'expression, le tout à l'aide d'un seul modèle et d'une seule branche.

- Nous démontrons que notre modèle produit des représentations 3D de visage de haute qualité, en utilisant une évaluation quantitative et qualitative, comparée aux méthodes de pointe dans le domaine.

1.2 Problématique et Objectifs de la Thèse

La problématique au cœur de cette thèse réside dans la nécessité impérieuse d'améliorer la reconstruction tridimensionnelle des visages pour répondre aux exigences croissantes des industries telles que l'animation, les jeux vidéo et l'esthétique. Les méthodes traditionnelles ont montré leurs limites, en particulier en ce qui concerne la capture précise des caractéristiques du visage dans des conditions variables. Par conséquent, cette recherche vise à développer une méthode innovante pour surmonter ces défis et contribuer de manière significative à l'évolution de la reconstruction 3D des visages.

Parmi les éléments de notre anatomie, le visage humain se distingue par sa complexité et son expressivité exceptionnelles. Il sert de vecteur clé pour la communication, la reconnaissance, et l'expression des émotions. Dans des domaines tels que le cinéma d'animation, les jeux vidéo et l'industrie esthétique, la demande de modèles 3D de visages précis est en constante augmentation. Les visages tridimensionnels réalistes sont essentiels pour créer des personnages convaincants, des avatars interactifs et des simulations d'intervention chirurgicale. Cependant, les approches traditionnelles de la reconstruction 3D des visages sont confrontées à des défis majeurs.

Les techniques classiques de reconstruction 3D des visages ont montré leurs limites, principalement en raison de leur incapacité à capturer avec précision les variations subtiles de l'expression faciale, les effets d'illumination, les angles de vue variables et les occultations partielles. Ces défis inhérents ont créé une disparité entre les attentes croissantes de l'industrie et les capacités actuelles de la technologie. Il devient donc impératif de développer des méthodes novatrices pour relever ces défis et perfectionner la reconstruction 3D des visages.

Cette thèse a pour objectifs de répondre à ces enjeux majeurs en proposant une approche basée sur les réseaux de neurones génératifs pour reconstruire des visages en 3D à partir d'une unique image en deux dimensions et des points de repère du visage. Les points de repère du visage serviront d'entrée au générateur, ce qui permettra de créer une représentation 3D du visage prenant en compte à la fois la géométrie et l'expression faciale, le tout dans un seul modèle intégré.

Le premier objectif consiste à concevoir et mettre en œuvre une méthodologie basée sur les réseaux de neurones génératifs pour reconstruire des visages en 3D. L'utilisation de cette architecture avancée permettra d'améliorer considérablement la précision de la représentation 3D du visage générée. Les points de repère du visage joueront un rôle central dans notre approche. En les utilisant comme entrée du générateur, nous visons à créer une représentation 3D du visage qui tienne compte à la fois de la géométrie et de l'expression faciale. Cette approche vise à résoudre certains des défis majeurs liés à la variabilité des visages humains. En effet, l'objectif ultime de cette recherche est d'évaluer les performances de notre modèle de reconstruction 3D des visages en comparaison avec les méthodes existantes dans le domaine. Cette évaluation se fera à la fois de manière quantitative, en mesurant des métriques précises, et qualitative, en évaluant la qualité visuelle des reconstructions.

1.3 Plan de la Thèse

Le plan de cette thèse est structuré en plusieurs étapes cruciales qui nous guideront vers l'atteinte de nos objectifs de recherche. Chaque chapitre contribue à notre compréhension de la reconstruction 3D des visages en utilisant des réseaux de neurones génératifs et des graph convolutional networks, tout en détaillant les aspects théoriques, méthodologiques, et d'évaluation de notre approche. Le plan de la thèse se présente comme suit :

Chapitre 2 : Apprentissage Profond Tridimensionnel Dans ce chapitre, nous explorons les bases de l'apprentissage profond appliqué à la modélisation tridimensionnelle. Nous commençons par aborder les différentes représentations des données 3D, qui sont essentielles pour la compréhension et la manipulation des objets tridimensionnels. Nous plongeons ensuite dans les concepts fondamentaux de l'apprentissage profond. Nous explorons également des concepts clés, ainsi les architectures de réseaux de neurones convolutionnels et les réseaux de convolution sur graphe et en mettant particulièrement l'accent sur les Réseaux Adversaires Génératifs.

Chapitre 3 : Travaux connexes sur la Reconstruction 3D des Visages Ce chapitre passera en revue les travaux antérieurs dans le domaine de la reconstruction 3D des visages. Nous mettrons en lumière les avancées récentes, les lacunes existantes et les opportunités pour notre approche. L'analyse de l'état de l'art servira de fondement pour notre propre travail de recherche.

Chapitre 4 : L'approche proposée Dans ce chapitre central, nous détaillerons l'architecture de notre générateur et de notre discriminateur. Nous expliquerons en profondeur comment nous utilisons les points de repère du visage comme entrée pour capturer la géométrie et l'expression faciale en 3D. La procédure d'entraînement de notre modèle sera présentée en détail, y compris les mécanismes d'optimisation utilisés pour obtenir des résultats de haute qualité.

Chapitre 5 : Évaluation et Expérimentations Ce chapitre sera consacré à l'évaluation approfondie de notre modèle. Nous évaluerons ses performances sous des critères quantitatifs et qualitatifs, en le comparant aux approches existantes. Les expérimentations nous permettront de valider l'efficacité de notre méthode et de mettre en lumière ses avantages et ses limites.

Chapitre 6 : Conclusion Générale Enfin, dans ce chapitre de conclusion, nous résumerons nos contributions majeures. Nous mettrons en évidence les avancées réalisées dans le domaine de la reconstruction 3D des visages grâce à notre approche. De plus, nous suggérerons des pistes de recherche futures pour continuer à améliorer les techniques de reconstruction 3D des visages en exploitant les réseaux de neurones génératifs et les graph convolutional networks.

Ce plan de thèse offre une structure logique pour aborder de manière approfondie la

reconstruction 3D des visages en utilisant des techniques d'apprentissage profond. Chaque chapitre contribue à notre compréhension globale du sujet et nous guide vers la réalisation de nos objectifs de recherche.

2

Apprentissage Profond Tridimensionnel

2.1 Introduction

Dans ce chapitre, nous examinerons les différentes méthodes de représentation des objets tridimensionnels, un aspect fondamental de la modélisation 3D. Une compréhension approfondie de ces représentations est essentielle pour notre travail sur la reconstruction 3D des visages en utilisant les modèles génératifs GAN. Nous explorerons les représentations basées sur des points 3D, des nuages de points, des maillages 3D, et des voxels 3D, en examinant leurs avantages, leurs inconvénients, et leurs applications potentielles dans notre contexte de recherche. Ensuite nous plongerons dans les bases de l'apprentissage profond, en nous concentrant sur les réseaux de neurones artificiels. Nous aborderons les concepts essentiels liés à l'apprentissage automatique, ainsi que l'architecture de base d'un réseau de neurones. Nous explorerons l'un des piliers de l'optimisation en apprentissage profond, à savoir la descente de gradient. Nous examinerons en détail son fonctionnement, son rôle dans l'ajustement des paramètres du modèle, et ses différentes variantes.

Nous entrerons dans le monde de la vision par ordinateur en explorant les réseaux de neurones convolutifs, nous étendrons notre compréhension en discutant des réseaux de convolution sur graphe, un domaine en plein essor dans l'apprentissage profond. Nous plongerons dans l'architecture des Réseaux Adversaires Génératifs, l'outil central de notre recherche pour reconstructruire des visages en 3D. Nous examinerons en détail comment les GAN fonctionnent, notamment leur générateur et leur discriminateur.

2.2 Représentation des Données 3D

2.2.1 Représentation des Points 3D

2.2.1.1 Coordonnées Cartésiennes

Les coordonnées cartésiennes sont la forme la plus courante de représentation des points 3D. Elles utilisent trois valeurs numériques (x, y, z) pour définir la position d'un point par rapport à un système de coordonnées orthogonal. Cette représentation est simple et intuitive, car elle correspond à notre perception naturelle de l'espace. Mathématiquement, un point 3D en coordonnées cartésiennes peut être exprimé comme suit :

$$P(x, y, z) \tag{2.1}$$

Où:
(x, y, z) sont les coordonnées du point par rapport à un référentiel fixe. Cette représentation

est largement utilisée dans les systèmes de rendu 3D, les jeux vidéo et la robotique.

2.2.1.2 Coordonnées Polaires

Les coordonnées polaires offrent une alternative intéressante aux coordonnées cartésiennes. Elles décrivent la position d'un point en utilisant la distance r par rapport à l'origine (appelée rayon) et l'angle θ entre la direction de référence et la ligne reliant l'origine au point. En 3D, une coordonnée polaire supplémentaire, ϕ , est ajoutée pour spécifier l'angle de rotation autour de l'axe initial. Mathématiquement, un point 3D en coordonnées polaires est défini comme suit :

$$P(r, \theta, \phi) \tag{2.2}$$

Où:

r est la distance du point à l'origine, θ est l'angle horizontal par rapport à l'axe de référence, et ϕ est l'angle vertical par rapport au plan horizontal. Les coordonnées polaires sont utiles dans des contextes où la direction et la distance par rapport à un point de référence sont plus pertinentes que les coordonnées cartésiennes. Par exemple, elles sont fréquemment utilisées en navigation aérienne.

2.2.1.3 Coordonnées Homogènes

Les coordonnées homogènes sont une représentation mathématique puissante des points 3D. Elles permettent de représenter des transformations géométriques (translation, rotation, mise à l'échelle) de manière élégante et efficace. Les coordonnées homogènes ajoutent une quatrième dimension w aux coordonnées cartésiennes classiques (x, y, z). Un point 3D en coordonnées homogènes est exprimé sous la forme :

$$P(x, y, z, w) \tag{2.3}$$

Où:
(x, y, z) sont les coordonnées cartésiennes et w est un facteur d'échelle. La transformation de ce point par une matrice de transformation 4×4 T est réalisée en effectuant le produit matriciel suivant :

$$P' = T \cdot P \tag{2.4}$$

Cette représentation est particulièrement utile dans la vision par ordinateur, la modélisation 3D et la réalité virtuelle, où des transformations complexes sont couramment appliquées.

En effet, La représentation des points 3D est un concept central dans de nombreuses disciplines, de la vision par ordinateur à la géométrie. Les coordonnées cartésiennes offrent une simplicité intuitive, tandis que les coordonnées polaires sont utiles pour des applications spécifiques où la direction et la distance sont cruciales. Les coordonnées homogènes, quant à elles, permettent d'effectuer des transformations géométriques de manière efficace. Le choix de la représentation dépend des besoins spécifiques de chaque application, et une compréhension approfondie de ces différentes méthodes est essentielle pour travailler efficacement dans le domaine de la représentation des points 3D.

2.2.2 Représentation en nuage de points

La représentation des données 3D en nuage de points est un domaine fondamental de la vision par ordinateur, de la robotique et de la modélisation 3D. Elle joue un rôle central dans la compréhension et la manipulation d'objets tridimensionnels dans un environnement numérique. La représentation des données 3D en nuage de points consiste à décrire un objet tridimensionnel en utilisant une collection de points dans l'espace 3D. Un exemple illustratif est présenté à la figure 2.1. Chaque point du nuage est généralement caractérisé par ses coordonnées spatiales (x, y, z) et éventuellement d'autres attributs tels que la couleur, la texture ou la densité. Cette représentation a une large gamme d'applications, allant de la reconstruction de scènes 3D à partir d'images 2D à la modélisation de la géométrie d'objets complexes.

La représentation des données 3D en nuage de points commence par la définition des coordonnées de chaque point dans l'espace 3D. Les coordonnées cartésiennes (x, y, z) sont couramment utilisées pour décrire la position d'un point dans un repère tridimensionnel. Ainsi, un nuage de points peut être représenté mathématiquement comme une collection de vecteurs :

$$P = \{(x_1, y_1, z_1), (x_2, y_2, z_2), \ldots, (x_n, y_n, z_n)\} \tag{2.5}$$

Où :
n est le nombre de points dans le nuage de points.

Outre les coordonnées spatiales, les nuages de points peuvent contenir divers attributs supplémentaires pour chaque point. Par exemple, dans les applications de numérisation laser, chaque point peut être associé à une intensité de réflexion lumineuse. Dans la numérisation 3D par photogrammétrie, la couleur de surface peut être enregistrée. Ces attributs supplémentaires enrichissent la représentation du nuage de points, permettant une description plus détaillée de l'objet ou de la scène.

Les données en nuage de points sont généralement acquises à l'aide de capteurs 3D, tels que des scanners laser, des caméras stéréo, des systèmes de projection de lumière structurée, ou même des drones équipés de capteurs 3D. Les capteurs 3D mesurent la distance entre le capteur et les points de la scène, ce qui permet de construire le nuage de points en temps réel ou a posteriori [3]. La représentation des données 3D en nuage de points est une composante essentielle de la vision par ordinateur, de la robotique et de la modélisation 3D. Elle permet de capturer la géométrie et d'autres attributs des objets tridimensionnels, ouvrant la voie à une multitude d'applications. Les avancées continues dans les capteurs 3D, les algorithmes de prétraitement et les méthodes de compression contribuent à rendre cette représentation plus efficace et précise que jamais, ouvrant la voie à de nouvelles possibilités dans de nombreux domaines.

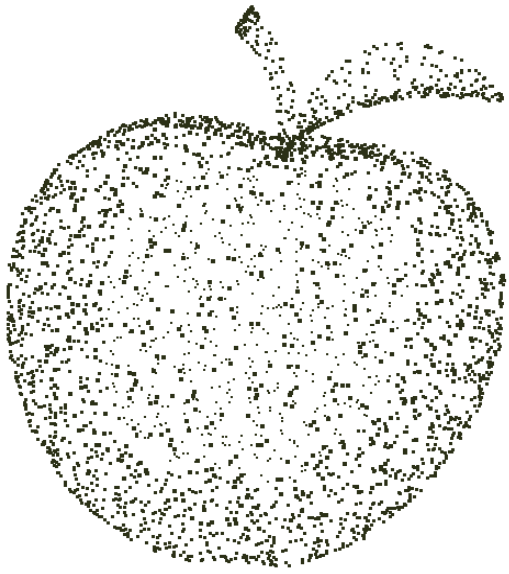

Figure 2.1: Exemple de représentation par nuage de points d'une pomme

2.2.3 Représentation maillée 3D

La représentation maillée 3D est une technique cruciale dans le domaine de la modélisation tridimensionnelle, utilisée pour représenter des objets, des scènes et des formes complexes en trois dimensions. Elle forme la base de nombreuses applications, de la conception assistée par ordinateur (CAO) à la simulation numérique, en passant par les jeux vidéo et la réalité virtuelle. Dans cette section, nous explorerons en détail la représentation maillée 3D, ses principes fondamentaux, ses avantages et ses limitations.

La représentation maillée 3D est basée sur la discrétisation de l'espace tridimensionnel en petits éléments appelés "mailles" ou "polygones". Un exemple illustratif est présenté à la figure 2.2. Ces mailles sont généralement des triangles (triangulation) ou des quadrilatères, bien que les triangles soient plus couramment utilisés en raison de leur simplicité et de leur efficacité. Mathématiquement, une maille est définie comme un ensemble de sommets connectés par des arêtes, formant une surface approximative de l'objet 3D. La représentation maillée 3D est essentiellement une approximation discrète d'une surface continue.

Un maillage 3D est généralement défini par trois composantes principales :

- **Sommets (Vertices)** : Chaque sommet est défini par ses coordonnées spatiales (x, y, z) dans l'espace tridimensionnel. Par exemple, un sommet peut être représenté comme. l'espace tridimensionnel. Par exemple, un sommet peut être représenté comme $V_i = (x_i, y_i, z_i)$

- **Arêtes (Edges)** : Les arêtes relient les sommets pour former les contours des mailles. Mathématiquement, une arête est définie par une paire de sommets connectés, par exemple, $E_{ij} = (V_i, V_j)$

- **Faces (Faces)** : Les faces sont les surfaces planes délimitées par les arêtes et les sommets. Dans le cas des triangles, chaque face est définie par trois sommets non alignés. Par exemple, une face peut être représentée comme $F_{ijk} = (V_i, V_j, V_k)$

La représentation maillée 3D présente de nombreux avantages qui en font un choix populaire pour la modélisation et la manipulation d'objets tridimensionnels. L'un des avantages majeurs est sa simplicité conceptuelle et de mise en œuvre. La structure maillée est intuitive et facile à comprendre, ce qui facilite son utilisation dans divers domaines.

Un autre avantage important est l'efficacité de stockage et de traitement. En utilisant des triangles, qui sont des structures géométriques simples, la représentation maillée permet de réduire considérablement la quantité de données nécessaires pour représenter une forme complexe par rapport à d'autres méthodes plus complexes, comme les nuages de points.

De plus, la représentation maillée est largement prise en charge par les moteurs graphiques modernes, ce qui en fait une option idéale pour la visualisation en temps réel, les simulations et les jeux vidéo. De plus, la plupart des logiciels de modélisation 3D et de rendu utilisent des maillages comme format d'entrée et de sortie standard.

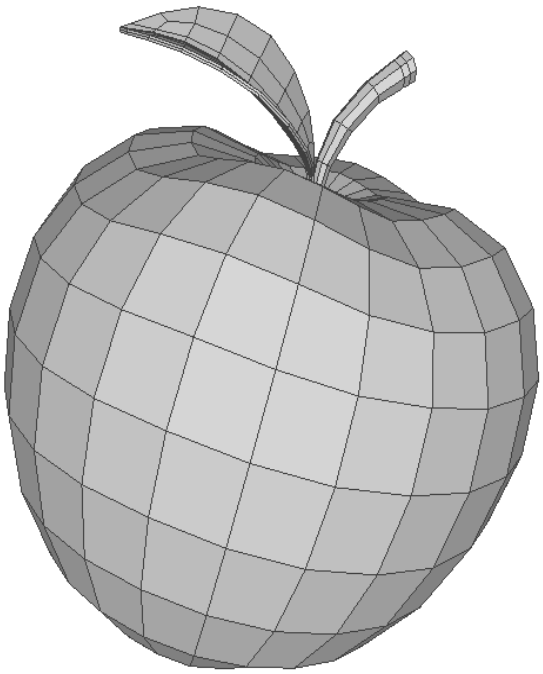

Figure 2.2: Représentation maillée 3D d'une pomme.

Cependant, la représentation maillée 3D présente également des limitations et des complexités importantes :

- Les maillages sont des approximations discrètes de surfaces continues. Plus la résolution du maillage est basse, plus l'approximation est grossière. Augmenter la résolution augmente la fidélité, mais cela entraîne également une augmentation de la complexité du modèle et de la charge de calcul associée.

- La topologie des maillages peut varier considérablement en fonction de la forme de l'objet. Les maillages peuvent avoir des trous, des déformations et des singularités, ce qui complique la manipulation et le traitement automatique.

- Les maillages haute résolution nécessitent une grande quantité de mémoire pour stocker les coordonnées de sommets, les arêtes et les faces. Cela peut poser des problèmes de performance dans les applications en temps réel.

- La création manuelle de maillages pour des objets complexes peut être une tâche fastidieuse et exigeante en temps. Des techniques de modélisation automatique et de simplification de maillage sont souvent nécessaires pour simplifier ce processus.

- L'application de textures sur des maillages peut être délicate en raison des distorsions qui peuvent survenir sur des surfaces courbes ou complexes.

2.2.4 Représentation des Voxels 3D

Une méthode courante pour représenter des objets 3D consiste à utiliser des voxels, qui sont l'équivalent tridimensionnel des pixels. La représentation des voxels 3D repose sur la discrétisation de l'espace tridimensionnel en petits éléments cubiques appelés voxels. Cette discrétisation permet de représenter un objet ou un volume tridimensionnel en attribuant des valeurs à chaque voxel, décrivant ainsi les propriétés de l'objet à cette position dans l'espace. Un volume 3D peut être représenté comme une grille régulière de voxels. Un exemple est présenté à la figure 2.3.

L'une des caractéristiques fondamentales des voxels est leur capacité à représenter des objets complexes avec une grande précision. Cette précision dépend de la résolution des voxels, c'est-à-dire de la taille de chaque élément cubique. Une résolution plus élevée permet une représentation plus précise, mais nécessite également une quantité de données plus importante.

Mathématiquement, une représentation en voxels 3D peut être décrite comme une fonction discrète définie sur une grille tridimensionnelle. Soit $V(x, y, z)$ la valeur du voxel à la position (x, y, z) , où x, y et z sont les coordonnées tridimensionnelles discrètes. Cette fonction $V(x, y, z)$ peut prendre diverses formes, en fonction de la nature de la donnée que

l'on souhaite représenter. Par exemple, dans le cas de données binaires, $V(x, y, z)$ peut être binaire, prenant les valeurs 0 ou 1 pour indiquer la présence ou l'absence d'un objet à la position (x, y, z) .

Lorsqu'il s'agit de données continues, chaque voxel peut contenir une valeur réelle représentant une caractéristique de l'objet à cette position. Ainsi, $V(x, y, z)$ peut être un nombre réel, tel que la densité, la température ou la concentration, en fonction de l'application. Dans le domaine de la vision par ordinateur, les représentations en voxels 3D sont utilisées pour la compréhension et la manipulation d'objets tridimensionnels dans des scènes. Les réseaux de neurones convolutifs tridimensionnels (3D CNN) sont couramment utilisés pour extraire des caractéristiques à partir de données voxelisées. Par exemple, dans la détection d'objets 3D, un modèle de réseau de neurones peut prendre en entrée une grille de voxels décrivant une scène 3D et générer des prédictions sur la présence et la position d'objets dans cette scène.

Les 3D Convolutional Neural Network (CNN) ont montré leur efficacité dans des tâches telles que la segmentation sémantique d'objets dans des nuages de points 3D, la reconnaissance d'objets dans des vidéos 3D et la reconstruction tridimensionnelle d'environnements à partir de capteurs LiDAR. Ils sont capables de capturer des informations complexes à partir de données voxelisées, mais leur utilisation efficace dépend de la conception du réseau, de la taille des données d'entraînement et de la qualité de la représentation voxelisée.

Les représentations en voxels 3D présentent plusieurs avantages significatifs. Tout d'abord, elles permettent de conserver une structure tridimensionnelle précise des objets, ce qui est essentiel dans de nombreuses applications, notamment la simulation numérique et la modélisation médicale. Ensuite, elles sont naturellement adaptées à la parallélisation, car chaque voxel peut être traité indépendamment, ce qui facilite leur manipulation en environnements de calcul haute performance. Enfin, elles sont intuitives et faciles à comprendre pour les êtres humains, ce qui les rend adaptées aux interfaces utilisateur tridimensionnelles.

Cependant, malgré leurs avantages, les représentations en voxels 3D présentent également des limitations. L'une des principales limitations réside dans leur consommation de mémoire. Plus la résolution des voxels est élevée, plus la quantité de mémoire requise est importante, ce qui peut devenir prohibitif pour les objets ou les scènes 3D complexes. En outre, les voxels souffrent souvent de ce qu'on appelle le "problème d'échelle", ce qui signifie que la même structure peut nécessiter un nombre variable de voxels en fonction de

sa taille, ce qui peut rendre difficile la comparaison entre différentes représentations.

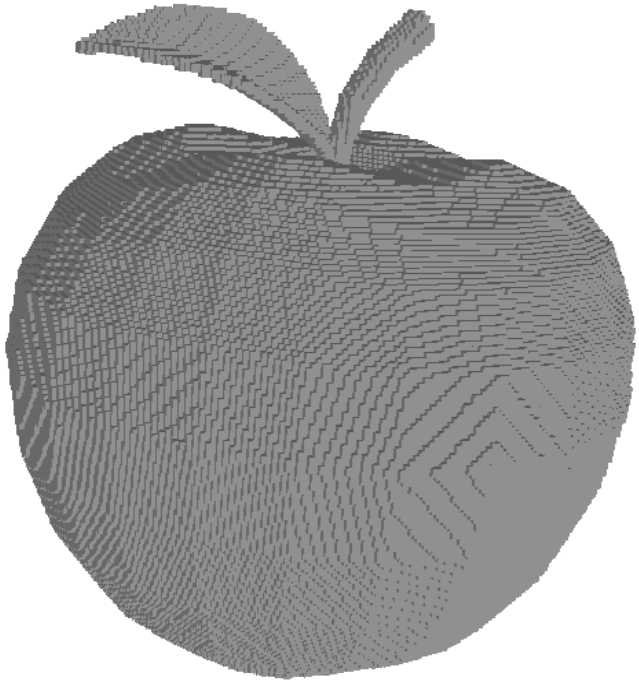

Figure 2.3: Exemple de représentation par Voxels d'une pomme.

2.3 Fondements théoriques de l'apprentissage profond

L'apprentissage profond, également connu sous le nom de "deep learning", est une sous-discipline de l'apprentissage automatique (machine learning) qui repose sur l'utilisation de réseaux de neurones artificiels profonds pour résoudre des tâches complexes d'analyse de données et de prise de décision. Cette approche tire son nom du fait qu'elle implique généralement la mise en œuvre de réseaux de neurones composés de plusieurs couches, également appelées couches cachées, permettant ainsi la création de modèles d'apprentissage capables de capturer des représentations hiérarchiques des données.

L'apprentissage profond vise à résoudre des problèmes en automatisant l'extraction de caractéristiques pertinentes des données brutes, ce qui le distingue des méthodes d'apprentissage plus traditionnelles où les caractéristiques doivent être définies manuellement par des experts. En d'autres termes, les réseaux de neurones profonds sont conçus pour apprendre de manière autonome les caractéristiques discriminantes des données à partir desquelles ils

peuvent prendre des décisions ou effectuer des prédictions.

Le principal avantage de l'apprentissage profond est son aptitude à apprendre des représentations de données à différents niveaux d'abstraction. Au lieu de nécessiter une ingénierie minutieuse des caractéristiques par des experts, les caractéristiques sont apprises automatiquement à partir des données brutes [4]. Un réseau de neurones profond est représenté comme une composition de fonctions. Pour un réseau de neurones à propagation avant, cela peut être formulé comme suit :

$$\begin{aligned} h^{(1)} &= f\left(W^{(1)}x + b^{(1)}\right) \\ h^{(2)} &= f\left(W^{(2)}h^{(1)} + b^{(2)}\right) \\ &\vdots \\ y &= f\left(W^{(n)}h^{(n-1)} + b^{(n)}\right) \end{aligned} \tag{2.6}$$

Où :
$W^{(i)}$ sont les poids de la couche i et $b^{(i)}$ sont les biais correspondants.

La profondeur d'un réseau de neurones, représentée par le nombre de couches cachées, est ce qui distingue l'apprentissage profond des approches plus superficielles. Cette profondeur accrue permet aux réseaux de neurones de capturer des caractéristiques abstraites et complexes dans les données, ce qui les rend adaptés à un large éventail d'applications, notamment la vision par ordinateur, le traitement du langage naturel, la reconnaissance vocale, la recommandation de contenu, et bien d'autres. En effet, l'apprentissage profond est une approche d'apprentissage automatique qui repose sur des réseaux de neurones artificiels profonds pour extraire automatiquement des caractéristiques pertinentes à partir des données brutes, ouvrant ainsi la voie à des avancées significatives dans la résolution de problèmes complexes.

2.3.1 Apprentissage d'un réseau de neurones

Les algorithmes d'apprentissage automatique, au cœur de nombreuses avancées technologiques, sont des mécanismes informatiques fascinants capables d'acquérir des connaissances à partir de données. Cependant, la notion d'apprentissage suscite des interrogations. En 1997, Tom Mitchell [5] a tenté de cerner cette notion complexe en proposant la définition suivante : "On dit qu'un programme informatique apprend de l'expérience E par rapport à une classe de tâches T et à une mesure de performance P, si sa performance aux tâches

dans T, mesurée par P, s'améliore avec l'expérience E."

Cette définition, soulève plusieurs questions essentielles. Tout d'abord, qu'entend-on par "expérience" ? L'expérience, au sens de Mitchell, représente l'ensemble des données et des situations auxquelles l'algorithme est exposé pour acquérir des connaissances. Elle peut englober une multitude de scénarios, allant de la classification d'images à la prédiction de séries temporelles, en passant par la recommandation de produits. En somme, l'expérience est la matière première à partir de laquelle l'apprentissage se déploie.

Ensuite, quelles sont les "tâches" dont il est question ? Les tâches, au sens de Mitchell, constituent un ensemble de problèmes ou de missions que l'algorithme doit résoudre ou accomplir. Par exemple, dans le domaine de la vision par ordinateur, une tâche pourrait être la classification d'images en catégories spécifiques telles que "chats" ou "chiens". Dans le contexte de la recherche d'informations, une tâche pourrait être de recommander des articles de presse pertinents à un utilisateur. Ainsi, les tâches représentent la finalité de l'apprentissage. Enfin, comment mesurer la performance ? La mesure de performance, désignée par P, constitue la manière d'évaluer la qualité des réponses ou des décisions prises par l'algorithme dans le cadre des tâches définies. Dans certaines situations, cela peut être la précision d'une classification, la réduction de l'erreur quadratique moyenne dans une régression, ou encore le taux de recommandations pertinentes dans un système de recommandation. La mesure de performance est cruciale car elle permet de quantifier l'efficacité de l'apprentissage.

Il est important de noter que l'univers de l'apprentissage automatique offre une incroyable variété d'expériences E, de tâches T, et de mesures de performance P. Les expériences peuvent varier depuis l'analyse de données génomiques complexes jusqu'à la prédiction de la météo. Les tâches peuvent s'étendre du traitement automatique du langage naturel à la détection d'anomalies dans les réseaux informatiques. Les mesures de performance peuvent être adaptées aux spécificités de chaque tâche, garantissant ainsi que l'apprentissage est en adéquation avec les objectifs visés.
Il est également important de noter que la définition de Mitchell ne prétend pas fournir une formalisation rigoureuse pour chaque aspect de l'apprentissage automatique. Au contraire, elle offre un cadre conceptuel qui peut être appliqué de manière flexible dans une multitude de contextes. Cette souplesse permet aux chercheurs et aux praticiens d'explorer et de concevoir des algorithmes d'apprentissage adaptés à des problèmes toujours plus diversifiés.

2.3.2 La Descente de Gradient

Au cœur de nombreux algorithmes d'apprentissage profond se trouve une méthode d'optimisation fondamentale : la descente de gradient.
La descente de gradient est un algorithme d'optimisation qui vise à minimiser une fonction de coût, souvent notée comme $J(\theta)$, en ajustant itérativement les paramètres θ du modèle. Cette fonction de coût mesure l'écart entre les prédictions du modèle et les valeurs réelles dans le jeu de données. L'objectif est de trouver les valeurs optimales de θ qui minimisent cette fonction de coût.
Mathématiquement, la descente de gradient consiste à mettre à jour les paramètres θ comme suit :

$$\theta_{\text{nouveau}} = \theta_{\text{actuel}} - \alpha \cdot \nabla \cdot \mathcal{J}\left(\theta_{\text{actuel}}\right) \tag{2.7}$$

- θ_{nouveau} représente les nouveaux paramètres après la mise à jour.
- θ_{actuel} représente les paramètres actuels.
- α est le taux d'apprentissage (learning rate), un hyperparamètre qui contrôle la taille des pas de mise à jour.
- $\nabla \cdot \mathcal{J}\left(\theta_{\text{actuel}}\right)$ est le gradient de la fonction de coût par rapport aux paramètres actuels.

2.3.3 Taux d'Apprentissage (Learning Rate)

Le taux d'apprentissage, souvent noté sous le symbole "α" ou "η" dans les algorithmes d'optimisation, est l'un des hyperparamètres cruciaux dans l'entraînement des réseaux de neurones en apprentissage profond. Il détermine la taille des pas que l'optimiseur prend lors de la descente de gradient, ce qui a un impact significatif sur la convergence de l'algorithme d'optimisation et, par conséquent, sur la performance du modèle. L'importance du taux d'apprentissage a été mise en évidence par de nombreuses études. Le taux d'apprentissage est le paramètre le plus critique à ajuster pour l'entraînement réussi d'un réseau de neurones profonds [6]. Le taux d'apprentissage intervient dans l'algorithme de descente de gradient, notamment dans la mise à jour des poids d'un modèle. La formule générale pour la mise à jour des poids "w" lors de la descente de gradient stochastique est donnée par :

$$w = w - \alpha * \nabla L(w) \tag{2.8}$$

où :

- "w" représente les poids du réseau de neurones,
- "α" est le taux d'apprentissage,
- "$\nabla L(w)$" est le gradient de la fonction de coût "L" par rapport aux poids "w".

La valeur du taux d'apprentissage "α" détermine la quantité par laquelle les poids sont ajustés à chaque itération de l'optimisation. Un taux d'apprentissage trop petit peut entraîner une convergence lente ou une convergence vers un minimum local, tandis qu'un taux d'apprentissage trop élevé peut entraîner une divergence de l'optimisation. Il existe plusieurs approches pour régler automatiquement le taux d'apprentissage au cours de l'entraînement. L'une des méthodes les plus couramment utilisées est l'optimisation adaptative, qui ajuste le taux d'apprentissage en fonction des caractéristiques locales de la fonction de coût. Un exemple bien connu de ces méthodes est l'algorithme "AdaGrad" [7], qui adapte le taux d'apprentissage pour chaque paramètre en fonction de son historique de gradients.

2.3.4 Convergence de la Descente de Gradient

La convergence de la descente de gradient est un sujet étudié en profondeur dans l'apprentissage machine. Comme le soulignent Bottou et al. (2018) [8] La descente de gradient est une méthode d'optimisation qui consiste à ajuster itérativement les paramètres d'un modèle pour minimiser une fonction de coût. Cette fonction de coût est souvent formulée comme une fonction convexe ou non convexe, et l'objectif est de trouver les valeurs de paramètres qui minimisent cette fonction. L'algorithme de descente de gradient met à jour les paramètres dans la direction opposée au gradient de la fonction de coût par rapport aux paramètres actuels. Cette direction est souvent appelée la "direction de descente", d'où le nom de l'algorithme.
La convergence de la descente de gradient se réfère à la propriété de l'algorithme qui garantit qu'il converge vers une solution optimale (c'est-à-dire, un minimum global ou local) à mesure que le nombre d'itérations augmente. Cette convergence est un élément essentiel de la garantie que la descente de gradient finira par trouver une solution acceptable, bien que cela puisse prendre plus ou moins de temps en fonction de divers facteurs.
Plusieurs théorèmes de convergence sont essentiels pour comprendre la convergence de la descente de gradient. L'un des plus fondamentaux est le théorème de convergence du gradient, qui stipule que si la fonction de coût est convexe et que le taux d'apprentissage (learning rate) est suffisamment petit, alors la séquence des itérés générés par la descente de gradient converge vers un minimum global de la fonction. La figure 2.4 a montré que si le taux d'apprentissage est élevé, la solution est ignorée, mais s'il est petit, il faut

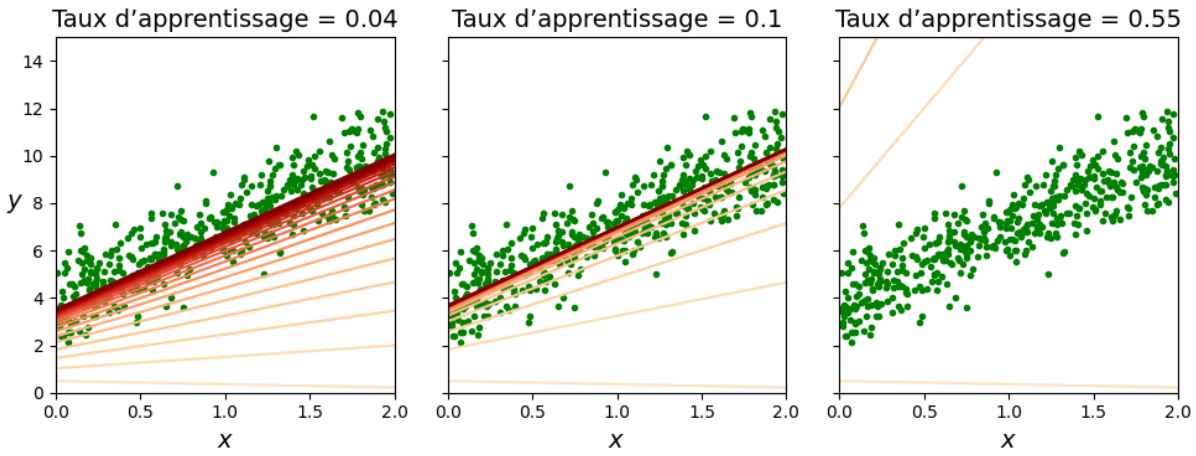

Figure 2.4: La convergence du descente de gradient.

beaucoup de temps pour la trouver. En revanche, avec un taux d'apprentissage moyen, la solution est atteinte dans un temps optimal.
Mathématiquement, cela peut être formulé comme suit pour une fonction de coût convexe $J(\theta)$ et un taux d'apprentissage.

$$\lim_{k \to \infty} \|\nabla J(\theta_k)\| = 0 \tag{2.9}$$

où θ_k est le vecteur de paramètres à l'itération k et $\nabla J(\theta_k)$ est le gradient de J évalué en θ_k. Un autre théorème important est le théorème de convergence sous des conditions plus générales. Il stipule que si la fonction de coût est continûment différentiable et que le taux d'apprentissage diminue suffisamment lentement, la séquence des itérés converge vers un minimum local.

2.3.5 Descente de Gradient Stochastique

Dans la Descente de Gradient Stochastique, le terme "stochastique" fait référence à l'utilisation d'un seul exemple de données à la fois pour mettre à jour les poids, contrairement à la descente de Gradient classique qui utilise l'ensemble des données d'entraînement à chaque itération. Cette approche stochastique présente plusieurs avantages significatifs. Permets les avantages del'algorithme SGD, l'efficacité Computationnelle et cela est due en utilisant un seul exemple à la fois. Le SGD peut être formulé mathématiquement comme suit :

$$\theta_{\text{nouveau}} = \theta_{\text{actuel}} - \alpha \cdot \nabla J(\theta_{\text{actuel}}; x_i, y_i) \tag{2.10}$$

Où le x_i est un exemple d'entraînement et y_i est la vérité terrain (la valeur réelle) associée à cet exemple.
La descente de gradient stochastique est beaucoup plus rapide, surtout lorsque l'ensemble de données est volumineux. Elle permet d'économiser des ressources computationnelles précieuses. Ainsi sa nature stochastique, lui permet d'échapper aux minima locaux plus facilement, car les mises à jour des poids sont moins déterministes. Elle a une meilleure capacité à explorer l'espace des paramètres. En revanche, l'inconvénient major de la méthode SGD est l'utilisation d'un seul exemple à la fois, les mises à jour des poids sont très variables d'une itération à l'autre, ce qui peut de rendre la convergence plus bruitée. Ainsi le choix d'un taux d'apprentissage trop élevé peut entraîner des oscillations autour du minimum, tandis qu'un taux d'apprentissage trop faible peut ralentir la convergence.

2.3.6 Descente de Gradient par mini lots

La descente de gradient par mini lot (batches) est une extension de la SGD qui cherche à combiner les avantages de la descente de gradient stochastique et de la Descente de Gradient classique. Au lieu d'utiliser un seul exemple ou l'ensemble complet des données, la Mini-batch SGD divise les données en petits ensembles appelés mini lots. Les mises à jour des poids sont calculées en moyennant les gradients des mini-batchs.

$$\theta = \theta - \alpha(1/b) \sum \nabla J(\theta) \tag{2.11}$$

Où b est la taille du mini-batch.
Permets les avantages de la descente de gradient par mini lots, l'équilibre entre efficacité et stabilité. En effet La la descente de gradient par mini lots offre un équilibre entre l'efficacité de la SGD et la stabilité de la descente de gradient classique. Elle est donc largement utilisée en pratique. Un autre point positif de cette dernière est le parallélisme. Les mini lots peuvent être calculés en parallèle, ce qui accélère encore le processus d'entraînement sur des architectures matérielles adaptées.

2.3.7 Réseaux de Neurones Convolutifs (CNN)

Les Réseaux de Neurones Convolutifs (CNN) représentent une avancée majeure dans le domaine de la vision par ordinateur et de l'apprentissage profond. Ces réseaux ont révolutionné la manière dont nous traitons et comprenons les données visuelles, en permettant des réalisations telles que la classification d'images, la détection d'objets et la segmentation sémantique. Inspirés par le fonctionnement du cortex visuel chez les animaux, les CNN sont conçus pour capturer efficacement les motifs et les caractéristiques des images, ce qui en fait des outils essentiels dans la compréhension des données visuelles complexes.

2.3.7.1 Architecture des Réseaux de Neurones Convolutifs

Les CNN se distinguent par leur architecture qui exploite les propriétés de la convolution, de la mise en commun (pooling) et de la non-linéarité.
La couche de convolution est au cœur de cette architecture et consiste en des filtres qui glissent sur l'image d'entrée, extrayant ainsi des caractéristiques locales. Chaque filtre effectue une opération de convolution en appliquant un produit scalaire entre les valeurs des pixels d'une région locale de l'image et les poids du filtre (voir la figure 2.5). Cette opération peut être formulée mathématiquement comme suit :

$$S(i,j) = (I * K)(i,j) = \sum_m \sum_n I(i+m, j+n) \cdot K(m,n) \tag{2.12}$$

Où $S(i,j)$ est la sortie du filtre à la position (i,j), I est l'image d'entrée, K est le noyau (filtre) et (m,n) sont les indices de la matrice du filtre.

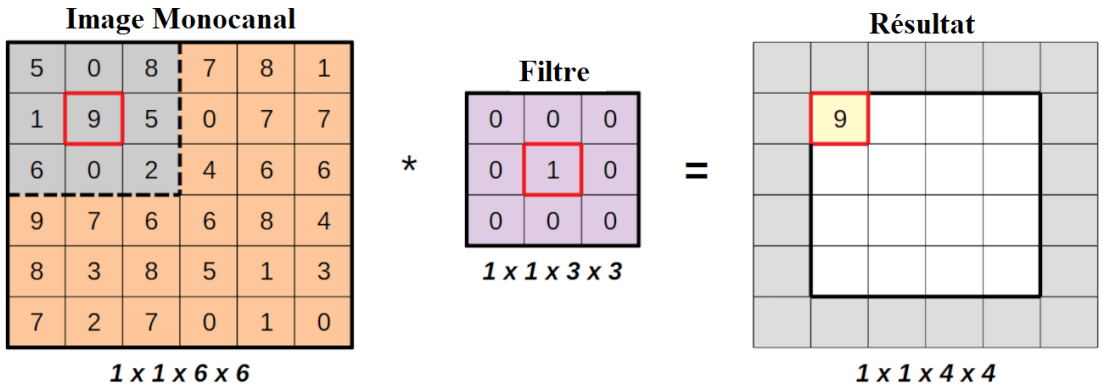

Figure 2.5: Les calculs primaires exécutés à chaque étape de la couche convolutive.

La couche de Pooling : Après les couches de convolution, les couches de mise en commun (pooling) sont utilisées pour réduire les dimensions spatiales des caractéristiques extraites, tout en préservant leur information essentielle. L'opération de mise en commun sélectionne la valeur maximale (max-pooling) ou moyenne (average-pooling) à partir d'une région locale de l'image. La figure 2.6 présente l'opération de max-pooling. Cela permet une invariance aux translations mineures dans l'image et contribue à la robustesse du réseau.

Les fonctions d'activation introduisent la non-linéarité dans les CNN, ce qui permet au réseau de capturer des relations complexes entre les caractéristiques. Une fonction d'activation couramment utilisée est la fonction Rectified Linear Unit (ReLU) (Rectified Linear Unit) définie comme :

$$f(x) = \max(0, x) \tag{2.13}$$

L'apprentissage des CNN se fait généralement par rétropropagation du gradient, où l'erreur entre les sorties prédites et les valeurs réelles est propagée en arrière à travers le réseau. L'objectif est de minimiser une fonction de coût à l'aide d'algorithmes d'optimisation tels que la descente de gradient stochastique. La mise en œuvre de cette optimisation nécessite l'ajustement des poids des connexions neuronales en fonction du gradient de

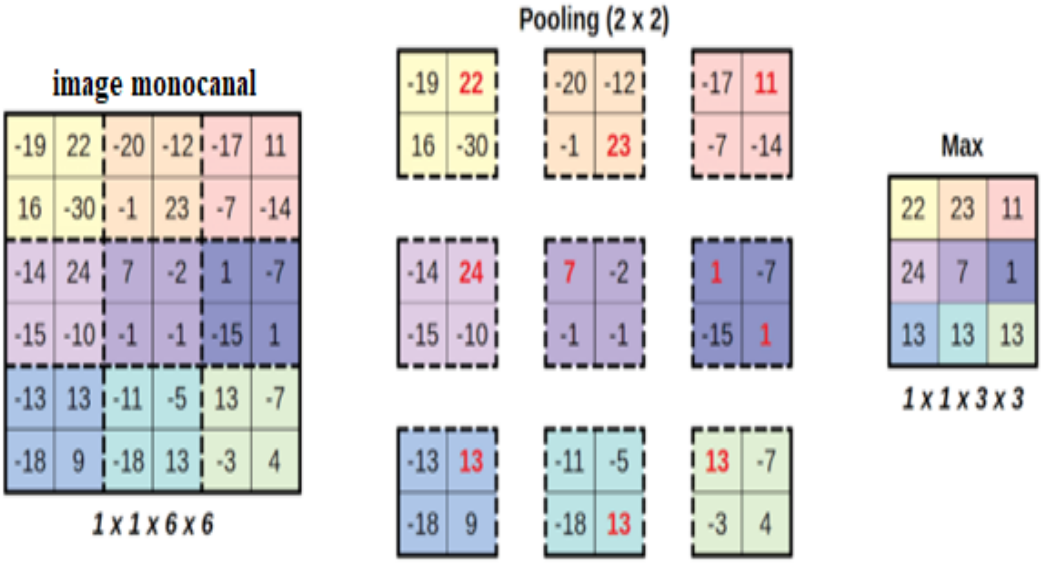

Figure 2.6: Exemple de couche de pooling max avec dimension 2×2

la fonction de coût par rapport à ces poids. Les CNN ont généré un impact significatif dans de nombreux domaines. Leur application à la classification d'images a conduit à des percées majeures, avec des modèles tels que AlexNet [9] et VGG [10] qui ont démontré des performances exceptionnelles dans la reconnaissance d'images. Les architectures plus récentes comme ResNet [11] ont introduit des blocs résiduels qui facilitent l'entraînement de réseaux plus profonds, tandis que les réseaux générateurs adverses [1] ont révolutionné la génération d'images réalistes.

2.3.8 Réseaux de Convolution sur Graphe (GCN)

Les Réseaux de Convolution sur Graphe (GCN, pour Graph Convolutional Networks) ont émergé comme une percée significative dans le domaine de l'apprentissage profond. Contrairement aux réseaux de convolution classiques qui excellent dans le traitement des données structurées régulières telles que les images, Les Graph Convolutional Network (GCN) ont été introduits par Thomas Kipf et Max Welling en 2016 [12], ouvrant ainsi la voie à une exploration plus profonde de la structure des données sur graphe et pour traiter des données non structurées et complexes, telles que les réseaux sociaux, les molécules chimiques et les systèmes de recommandation.

Les GCN ont été introduits pour appliquer les opérations de convolution aux données de graphe. Contrairement aux CNN, où la convolution est appliquée à des données régulières

(comme des matrices d'images), les GCN s'adaptent à la topologie variable des graphes. Les opérations de convolution sur les graphes permettent d'extraire des caractéristiques importantes des nœuds en tenant compte de leurs voisins directs. Plusieurs contributions ont été apportées afin de résoudre le problème de localisation spéciale des données [13]–[16].

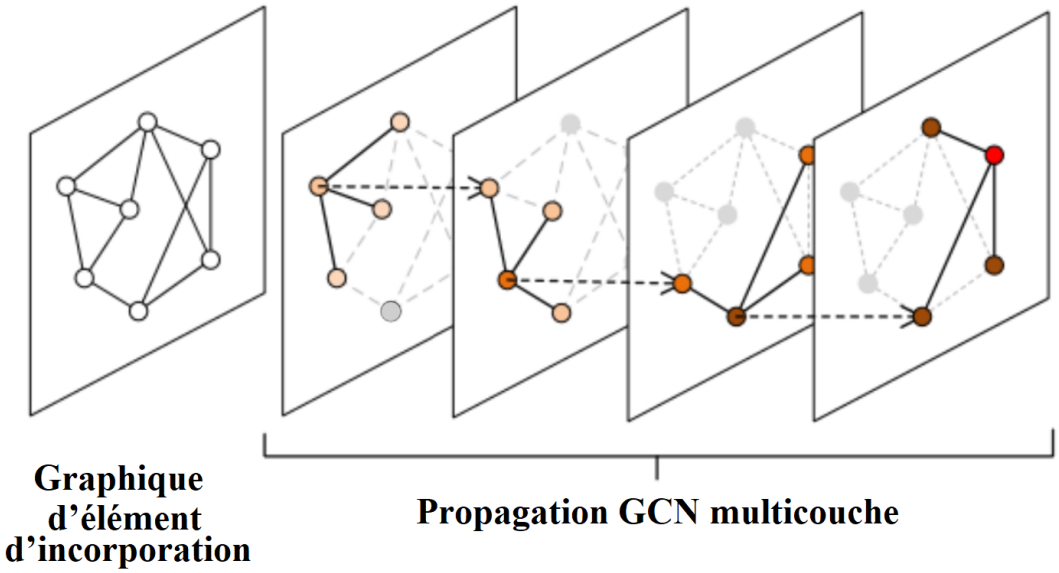

Figure 2.7: Illustration du processus de propagation du GCN [17].

Les GCN s'appuient sur des principes mathématiques solides pour effectuer des opérations de convolution sur des graphes. L'idée centrale est de propager l'information des nœuds voisins vers un nœud donné tout en tenant compte de la structure du graphe. Cela est accompli en utilisant la notion de matrice d'adjacence, qui représente les relations entre les nœuds.

Soit un graphe non dirigé représenté par une paire (V, E), où Vest l'ensemble des nœuds et E est l'ensemble des arêtes. Chaque nœud $v_i \in V$ est associé à une caractéristique initiale x_i. La figure 2.8 illustre une représentation du graphe.

L'objectif des GCN est de mettre à jour ces caractéristiques en fonction des caractéristiques de leurs voisins. La mise à jour des caractéristiques d'un nœud v_i à une couche donnée est donnée par :

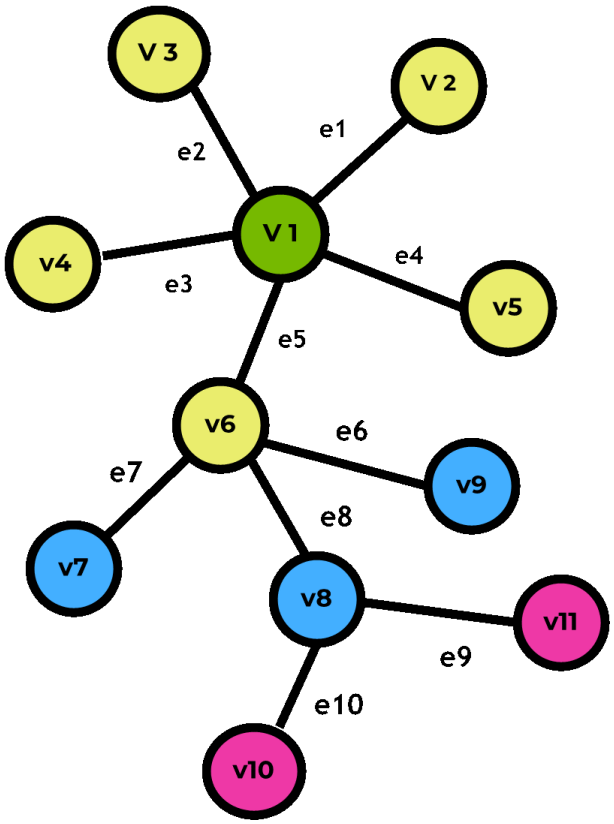

Figure 2.8: Illustration d'une représentation du graphe.

$$f_p^{l+1} = w_0 f_p^l + \sum_{q \in \mathcal{N}(p)} w_1 f_q^l \tag{2.14}$$

Où :

- $f_p^{(l)}$ est la caractéristique du nœud p à la couche l.
- $N\,(p)$ est l'ensemble des voisins du nœud p.
- w_0 et w_1 sont les poids à la couche l.

La propagation de l'information dans les GCN est itérative, chaque couche mettant à jour les caractéristiques des nœuds en fonction de celles de leurs voisins. Cette propagation continue jusqu'à ce que les caractéristiques convergent vers une représentation finale. L'image 2.7 illustre le processus de propagation du GCN.

2.3.8.1 Avantages des Réseaux de Convolution sur Graphe

Modélisation des Relations Complexes

Les Réseaux de Convolution sur Graphe offrent plusieurs avantages significatifs, notamment la capacité de modéliser des relations complexes et non linéaires entre les entités dans un graphe. Cette capacité est essentielle pour de nombreuses applications du monde réel.

Adaptabilité aux Données de Graphe

Contrairement aux CNN, qui supposent souvent des données structurées régulières comme les images, les GCN sont conçus pour être flexibles et s'adapter à la topologie variable des graphes. Cela signifie qu'ils peuvent être appliqués à une gamme diversifiée de problèmes, allant de la détection de fraude dans les transactions financières à la prédiction de la structure des protéines.

Transfert de Connaissances

Les GCN sont également efficaces pour le transfert de connaissances d'un domaine à un autre. Une fois qu'un GCN est pré-entraîné sur un graphe, ses représentations apprises peuvent être transférées vers un autre graphe pour une tâche similaire. Cela permet de tirer parti de l'expérience acquise sur un ensemble de données pour améliorer les performances sur un autre ensemble de données.

2.3.8.2 Limitations des Réseaux de Convolution sur Graphe

Malgré leurs avantages, les GCN présentent également certaines limitations et défis importants.

Besoin de Données Étendues

Comme de nombreux modèles d'apprentissage profond, les GCN nécessitent souvent une grande quantité de données étiquetées pour s'entraîner efficacement. Dans de nombreux domaines. De plus, l'étiquetage de données sur un graphe peut être complexe, car il nécessite souvent des annotations spécifiques à la structure du graphe.

Sensibilité à la Topologie du Graphe

Les performances des GCN dépendent largement de la topologie du graphe sous-jacent. Les graphes non informatifs, mal connectés ou mal équilibrés peuvent poser des défis. Les GCN ont du mal à gérer les nœuds avec degrés faibles ou élevés, ce qui peut entraîner des

biais dans les représentations apprises.

2.4 Les modèles génératifs

Les Réseaux adversaires génératifs ont révolutionné le domaine de l'apprentissage profond depuis leur introduction par Ian Goodfellow et ses collègues en 2014 [1]. Les GANs ont ouvert de nouvelles perspectives dans de nombreux domaines, notamment la génération d'images [18], la super-résolution d'images [19], la traduction automatique [20] et bien d'autres. Cette section plongera dans les fondements théoriques et les applications pratiques des GANs, en se penchant sur la manière dont ils exploitent des compétitions adverses pour générer des données réalistes.

Les GANs sont un cadre d'apprentissage profond composé de deux réseaux neuronaux, un générateur et un discriminateur, comme illustré à la figure 2.9. Ces deux réseaux sont en compétition constante. L'objectif principal des GANs est de générer des données réalistes à partir d'une distribution latente donnée. Le générateur tente de produire des données indiscernables des données réelles, tandis que le discriminateur cherche à distinguer les données générées de celles réelles. Cette compétition crée un équilibre dynamique où le générateur s'améliore constamment pour tromper le discriminateur, tandis que le discriminateur s'adapte pour devenir plus efficace dans la détection des faux.

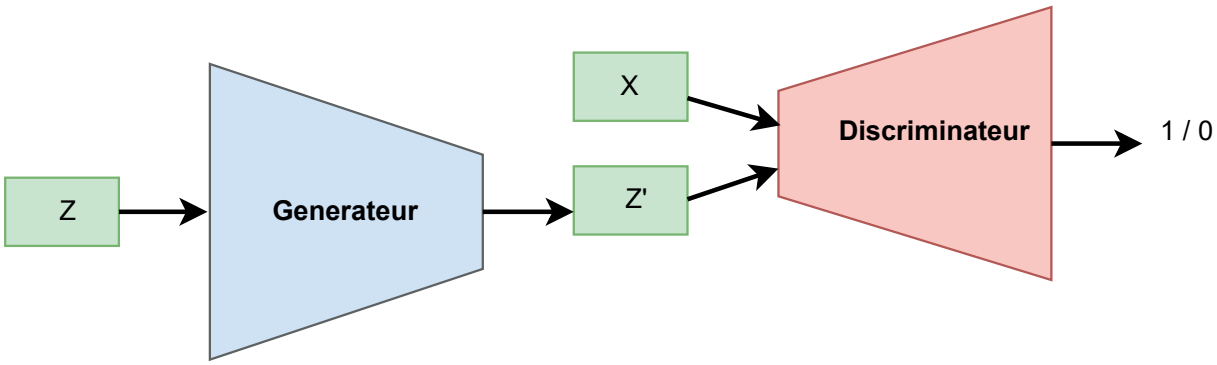

Figure 2.9: Architecture du réseau contradictoire génératif (GAN)

Mathématiquement, un GAN est défini comme un jeu à somme nulle où la fonction de perte du générateur et celle du discriminateur sont en opposition. La fonction de perte du générateur, notée L_G, est généralement basée sur la capacité du générateur à tromper le

discriminateur. La fonction de perte du discriminateur, notée L_D, quant à elle, mesure la capacité du discriminateur à différencier les données générées de celles réelles. Le but ultime est d'atteindre un équilibre où L_G et L_D sont minimales.

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{\text{data}}\,(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))] \tag{2.15}$$

Où :

- $P_{data(x)}$ représente la distribution des données réelles.
- z est la distribution latente.
- $G(z)$ est la sortie du générateur pour une entrée z.
- $D(x)$ est la sortie du discriminateur pour une entrée x.

L'objectif est de trouver le générateur G qui minimise cette fonction de perte tout en maximisant D pour créer des données générées de haute qualité.

2.4.1 Architecture des Réseaux Adversaires Génératifs

2.4.1.1 Le Générateur

Le générateur est l'un des composants essentiels dans les réseaux générateurs adverses. Les GANs sont conçus pour générer des données réalistes à partir de données de bruit aléatoire. Le générateur est chargé de cette tâche cruciale, créant des échantillons qui ressemblent autant que possible à des données réelles.

Le générateur peut être conceptualisé comme une fonction mathématique, généralement notée G, qui prend un vecteur de bruit aléatoire z comme entrée et produit une sortie qui ressemble à des données réelles. Mathématiquement, cela peut être exprimé comme suit :

$$\text{Générateur : } G(z) \to x \tag{2.16}$$

Où :

- G est la fonction générateur.
- z est le vecteur de bruit aléatoire (souvent tiré d'une distribution comme une distribution gaussienne).
- x est la sortie générée, qui est l'objet de référence pour ressembler à des données réelles.

Le générateur est généralement implémenté sous la forme d'un réseau de neurones, plus précisément, d'un réseau de neurones génératif. Ce réseau est conçu pour prendre en

entrée le vecteur de bruit z et produire une sortie qui peut être transformée en une image, un son, ou toute autre forme de données que le GAN est chargé de générer.

Une caractéristique essentielle du générateur est qu'il apprend à partir des données réelles disponibles lors de l'entraînement. Il ajuste ses paramètres de manière itérative pour minimiser la différence entre la sortie générée et les données réelles. Cela se fait en utilisant une fonction de coût, souvent une fonction de perte, qui mesure la divergence entre la distribution des données réelles et celle des données générées. Une fonction de coût couramment utilisée est la divergence de Kullback-Leibler (KL) ou la perte de log-vraisemblance négative.

$$\mathcal{L}_G = -\mathbb{E}_{z\sim p_z(z)}[\log(D(G(z)))] \tag{2.17}$$

Où :

- L_G est la fonction de coût du générateur.
- $p_z(z)$ est la distribution du bruit aléatoire.
- $D(G(z))$ est la sortie du discriminateur pour les données générées.

La tâche du générateur est de minimiser cette fonction de coût en ajustant ses paramètres de manière à tromper le discriminateur. En d'autres termes, il cherche à générer des données qui sont indiscernables des données réelles, de sorte que le discriminateur soit incapable de les distinguer.

2.4.1.2 Le Discriminateur

Le discriminateur, comme son nom l'indique, est le composant du GAN qui tente de discriminer entre les données réelles et les données générées par le générateur. Il est également implémenté sous la forme d'un réseau de neurones, mais contrairement au générateur, il est chargé de résoudre un problème de classification binaire : il doit décider si une donnée est réelle ou générée. Mathématiquement, cela peut être exprimé comme suit :

$$\text{Discriminateur} : D(x) \rightarrow [0, 1] \tag{2.18}$$

Où :

- D est la fonction discriminateur.

- x est une donnée en entrée (soit réelle, soit générée).
- $D(x)$ est la sortie du discriminateur, qui donne une probabilité que x soit une donnée réelle.

Le discriminateur est formé de manière à maximiser sa capacité à distinguer les données réelles des données générées. Cela revient à minimiser la fonction de coût du discriminateur, qui est souvent une fonction de perte de type binaire :

$$\mathcal{L}_D = -\left(\mathbb{E}_{x\sim p_{\text{data}}\,(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))]\right) \tag{2.19}$$

Où :

- L_D est la fonction de coût du discriminateur.
- $p_{data}(x)$ est la distribution des données réelles.
- $D(x)$ est la sortie du discriminateur pour les données réelles.
- $p_z(z)$ est la distribution du bruit aléatoire.
- $D(G(z))$ est la sortie du discriminateur pour les données générées.

L'objectif du discriminateur est donc d'attribuer des probabilités proches de 1 aux données réelles et des probabilités proches de 0 aux données générées.

L'architecture des GAN peut être modifiée et adaptée en fonction de l'application spécifique. Cependant, il existe deux architectures de base largement utilisées. Le GAN standard (simple), dans ce type, le générateur est généralement un réseau de neurones à propagation avant et le discriminateur est également un réseau de neurones à propagation avant. Ces deux réseaux sont formés de manière concurrente.

Les Deep Convolutional Generative Adversarial Network (DCGAN) utilisent des couches de convolution à la place des couches entièrement connectées, ce qui permet de générer des images de manière plus efficace. Cette architecture est couramment utilisée dans la génération d'images réalistes.

Il est important de noter que les GANs ne sont pas exempts de défis, notamment la convergence instable, le mode collapse et la génération de données biaisées. De nombreuses variantes des GAN ont été développées pour remédier à ces problèmes, notamment les Wasserstein Generative Adversarial Network (WGAN) [21] et les Conditional Generative Adversarial Network (CGAN) [22] ...etc.

2.4.2 Entraînement des Réseaux adversaires génératifs

L'entraînement des GANs est une étape cruciale et délicate. Le processus d'entraînement est itératif et consiste en l'optimisation concurrente du générateur et du discriminateur. Le processus d'entraînement des GAN consiste en une série d'itérations où le générateur et le discriminateur sont mis à jour de manière alternative pour atteindre un équilibre de Nash [23]. Lorsque cet équilibre est atteint, le générateur est capable de créer des données réalistes indiscernables des données réelles.

En effet, Le générateur est formé à minimiser la fonction de perte L_G , tandis que le discriminateur est formé à minimiser la fonction de perte L_D. Cette compétition constante entre les deux réseaux est ce qui conduit à l'amélioration continue de leurs performances respectives.

L'algorithme de rétropropagation (backpropagation) est utilisé pour mettre à jour les poids des réseaux. Le générateur ajuste ses poids pour maximiser la probabilité que le discriminateur commette une erreur en classifiant les données générées comme réelles. Inversement, le discriminateur ajuste ses poids pour minimiser l'erreur de classification entre les données réelles et générées. Veuillez consulter la figure 2.10 pour plus de détails.

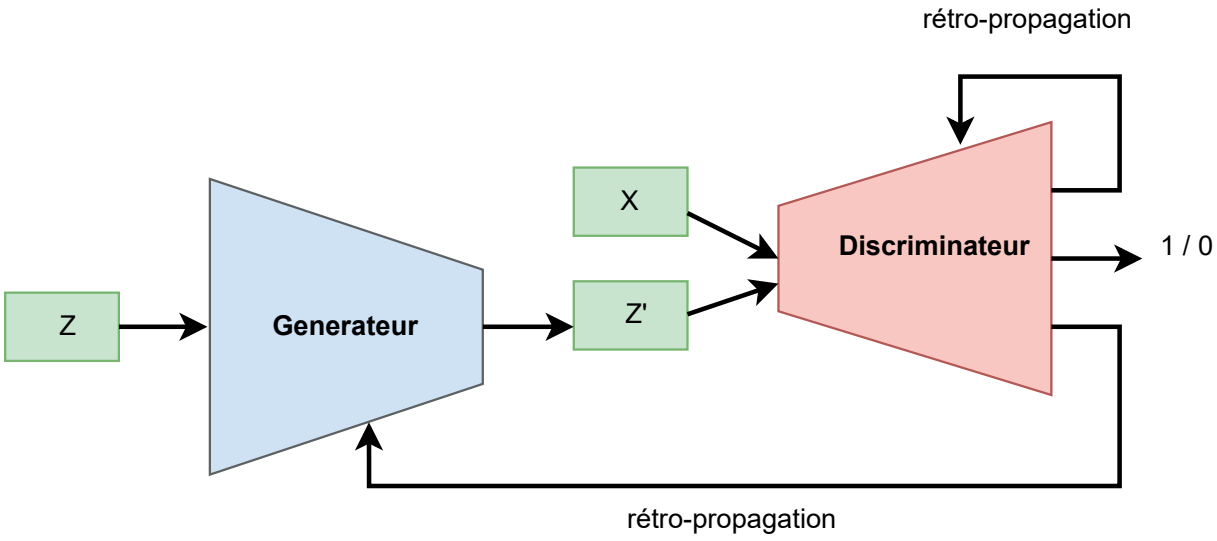

Figure 2.10: La rétropropagation d'un réseau contradictoire génératif (GAN)

Le processus d'entraînement des GANs peut être instable et sujet à des problèmes tels que l'effondrement du mode (mode collapse) où le générateur produit des données très similaires. Pour atténuer ces problèmes, des techniques avancées de régularisation comme le dropout dans le générateur, l'utilisation de la normalisation de lot (batch normalization), et l'optimisation des hyperparamètres sont souvent nécessaires.

2.4.3 Limitations et Défis des Réseaux adversaires génératifs

Bien que les GANs aient révolutionné de nombreux domaines, ils présentent également des limitations et des défis importants :

Mode Collapse : Lorsque le générateur ne produit qu'une petite gamme de données similaires, il est dit avoir subi un effondrement du mode. Cela limite la diversité des données générées. Sensibilité aux hyperparamètres : Les GANs sont sensibles aux choix des hyperparamètres, ce qui rend leur entraînement délicat et exigeant en termes de ressources.

Interprétabilité limitée : Il est souvent difficile de comprendre pourquoi un GAN génère une certaine sortie, ce qui rend leur interprétation problématique. Besoin de données massives : Les GANs ont besoin de grandes quantités de données pour fonctionner correctement, ce qui peut être un obstacle dans les domaines où les données sont rares.

Éthique et biais : L'utilisation de GANs soulève des préoccupations éthiques, notamment en ce qui concerne la création de fausses informations et la propagation de biais existants dans les données d'entraînement.

2.5 Conclusion

Ce chapitre a posé les bases conceptuelles nécessaires pour la compréhension de notre recherche en reconstruction 3D à l'aide de modèles génératifs. nous avons établi les fondements théoriques essentiels qui sous-tendent notre recherche en reconstruction tridimensionnelle à l'aide de techniques d'apprentissage profond, en mettant en lumière les différents aspects nécessaires à la compréhension de notre approche. nous avons exploré les diverses représentations des données 3D, en mettant l'accent sur les points 3D, les nuages de points, les maillages 3D et les voxels 3D. Comprendre ces différentes formes de représen-

tation est essentiel pour l'interprétation et la manipulation des données tridimensionnelles.

Nous avons également plongé dans les fondements théoriques de l'apprentissage profond, en commençant par l'apprentissage des réseaux de neurones. Cette section a abordé des concepts clés tels que les fonctions d'activation, la rétropropagation et les architectures de réseaux.

Nous avons examiné en détail l'optimisation en apprentissage profond, en mettant l'accent sur des éléments essentiels tels que la descente de gradient, le taux d' apprentissage, la convergence et l'utilisation de mini-lots pour l'entraînement efficace des réseaux de neurones. Les réseaux de neurones convolutifs ont également été présentés, soulignant leur importance dans le traitement d'images et la reconnaissance de motifs. Nous avons également exploré les réseaux de convolution sur graphe, qui élargissent les capacités des CNN pour la représentation et l'analyse de données complexes basées sur des graphes. Enfin, nous avons introduit les modèles génératifs, en particulier l'architecture des Réseaux Adversaires Génératifs. Nous avons examiné le processus d'entraînement des GAN et identifié certaines des limitations et des défis associés à ces modèles.

Dans le prochain chapitre, nous plongerons dans l'état de l'art actuel en matière de reconstruction 3D, en mettant en évidence les travaux pertinents dans ce domaine en constante évolution. Ces connaissances préparatoires seront cruciales pour la formulation de notre propre approche innovante.

3

Travaux connexes sur la Reconstruction 3D des Visages

3.1 Introduction

Ce chapitre est dédié à l'exploration approfondie de l'état de l'art en matière de reconstruction tridimensionnelle. Ce domaine de recherche joue un rôle fondamental dans des applications variées allant de la réalité virtuelle à la médecine, en passant par la vision par ordinateur et l'industrie du divertissement. Au cours de ce chapitre, nous aborderons diverses approches et techniques qui ont été développées pour résoudre le problème complexe de la reconstruction 3D à partir de données en deux dimensions (2D).

Nous explorerons les avancées majeures dans le domaine de la reconstruction tridimensionnelle à partir d'images. Nous étudierons les méthodes qui ont été développées pour capturer la troisième dimension à partir de données 2D, notamment à travers l'utilisation de modèles déformables tridimensionnels. Ces modèles, qui ont évolué au fil des années, constituent l'une des approches fondamentales pour résoudre ce problème complexe. Nous examinerons également les techniques basées sur la structure acquise à partir du mouvement (SfM), qui ont révolutionné la reconstruction 3D en exploitant les mouvements relatifs entre les objets et la caméra.

La deuxième section de ce chapitre se concentrera spécifiquement sur la reconstruction tridimensionnelle du visage, une tâche cruciale dans des domaines tels que la réalité augmentée, l'animation numérique et la sécurité biométrique. Nous aborderons en détail les modèles antagonistes (GAN) et leur application à la reconstruction 3D du visage. Ces modèles ont récemment suscité un grand intérêt en raison de leur capacité à générer des représentations tridimensionnelles réalistes à partir de données en 2D. Nous examinerons les avancées significatives réalisées dans ce domaine et évaluerons les défis qui subsistent.

3.2 Reconstruction 3D à partir des images

3.2.1 Modèles déformables tridimensionnelle

La modélisation déformable tridimensionnelle, également connue sous le nom de 3D Morphable Model (3DMM), est une technique avancée utilisée pour représenter et modéliser des objets ou des surfaces tridimensionnelles, tels que des visages humains, des corps, des objets, etc. Les modèles déformables tridimensionnels sont largement utilisés dans des domaines tels que la vision par ordinateur, la réalité virtuelle, la réalité augmentée, la reconstruction 3D, l'animation, et bien d'autres. Ils permettent de capturer et de

représenter efficacement la variabilité complexe des formes et des déformations dans un espace de paramètres contrôlables, ce qui facilite de nombreuses applications pratiques.

Dans cet état de l'art, nous allons explorer plusieurs méthodes et approches liées aux modèles déformables tridimensionnels, en mettant l'accent sur les 3DMM appliqués à la modélisation des visages humains. Blanz et al. (1999) [24] Cette méthode est l'une des premières propositions de modèles déformables tridimensionnels pour la modélisation des visages humains, comme illustré dans la figure 3.1 qui présente les résultats générés avec ce modèle. Son principe consiste à développer un modèle statistique déformable qui permette de générer des visages humains réalistes en 3D à partir d'un ensemble d'exemples d'images de visages. Le modèle se base sur la décomposition en valeurs singulières (SVD) d'un ensemble de formes et de textures de visages, permettant ainsi de capturer efficacement la variabilité complexe des visages humains dans un espace de paramètres contrôlables. En effet, les auteurs ont rassemblé un ensemble de visages tridimensionnels en utilisant un scanner 3D pour capturer la géométrie du visage et une caméra couleur pour enregistrer les textures. Pour comparer les visages, il est nécessaire de les aligner dans une position et une échelle commune. Les auteurs ont utilisé un ensemble de repères faciaux pour aligner les visages 3D. Les formes et les textures des visages alignés sont décomposées en valeurs singulières, permettant de définir un espace de variation. Cela permet de créer une base de formes et de textures, appelée base de modèle, qui représente les principales variations observées dans l'ensemble de données.
Une fois le modèle construit, il est possible de reconstruire un visage en combinant la forme moyenne du visage avec les déformations correspondant aux coefficients de la base de modèle. Le modèle déformable ainsi créé peut être utilisé pour la synthèse de visages réalistes en 3D. En ajustant les coefficients de déformation, il est possible de générer des visages qui ressemblent à de nouvelles personnes ou qui présentent différentes expressions et poses [25]. Les auteurs présentent le "Basel Face Model" (BFM), un modèle génératif de forme et de texture 3D de visage. Ils ont amélioré les modèles précédents en offrant une meilleure précision de la forme et de la texture grâce à un meilleur dispositif de numérisation et moins d'artefacts de correspondance grâce à un algorithme d'enregistrement amélioré.

Le BFM peut être ajusté aux images 2D ou 3D acquises dans différentes situations et avec différents capteurs en utilisant une méthode d'analyse par synthèse. Les paramètres du modèle résultant séparent la pose, l'éclairage, l'imagerie et les paramètres d'identité, ce qui facilite la reconnaissance de visages invariants entre les capteurs et les ensembles de données en comparant uniquement les paramètres d'identité. Au-delà des visages. Des modèles ont

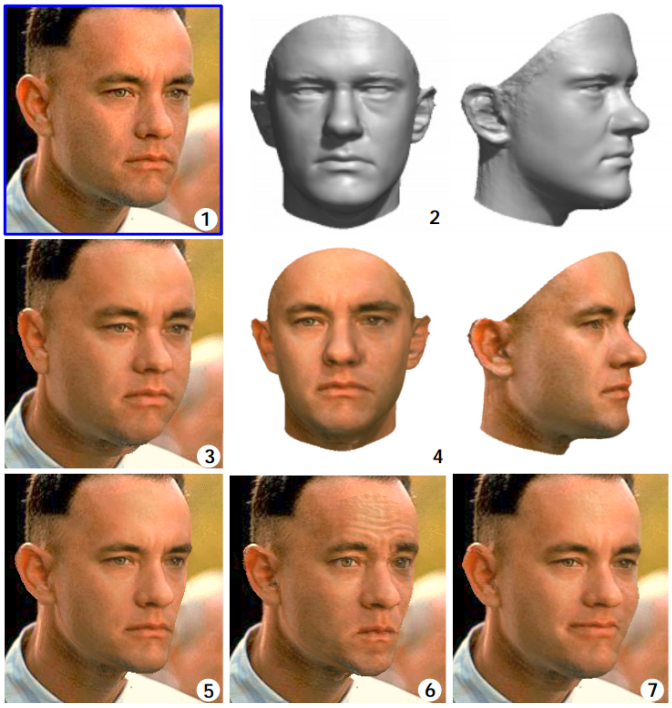

Figure 3.1: Résultats générés par le model modèle [24] une seule image en entrée (1) forme 3D produite (2) une estimation de la carte de texture (4) Le modèle 3D est ensuite rendu dans l'image après avoir modifié des caractéristiques faciales, telles que la prise de poids (3) et la perte de poids (5), le renfrogné (6) ou le sourire forcé (7).

été proposés pour la modélisation d'objets et de certaines parties du corps humain [26], [27].

Dans ce travail, Sahasrabudhe [28] introduisent une méthode non supervisée pour soulever une catégorie d'objets dans une représentation 3D, permettant d'apprendre un modèle 3D morphable de visages à partir d'un ensemble d'images non organisée. Cette approche ouvre de nombreuses perspectives pour la génération et la manipulation réaliste d'images en 3D, en particulier pour les visages humains. Elle permet d'obtenir des résultats convaincants en utilisant uniquement des informations non supervisées.

Tewari et al. [29] font partie des premiers dans leur tentative d'apprendre l'ajustement 3DMM à partir d'images non étiquetées. Ils utilisent une approche de perte non supervisée qui compare la texture projetée du maillage facial avec l'image originale elle-même. De plus, un alignement de points de repère clairsemé est utilisé comme perte auxiliaire. Gênes et al. [30] ont ensuite amélioré cette approche en comparant les images reconstruites avec l'entrée d'origine, en utilisant des caractéristiques de niveau supérieur issues d'un réseau

de reconnaissance faciale pré-entraîné. Contrairement à ces travaux, notre étude poursuit un objectif différent en cherchant à apprendre un 3DMM non linéaire.

3.2.2 Structure acquise à partir du mouvement (SfM)

Huber, P. [31] présente un modèle déformable tridimensionnel à résolution multiple pour la modélisation et l'ajustement des visages humains en 3D. Le modèle, appelé "Multiresolution 3D Morphable Face Model" (M3DFM), est conçu pour capturer la variabilité complexe des formes faciales et des expressions, permettant ainsi la génération réaliste de visages et l'ajustement précis à partir d'images 2D.

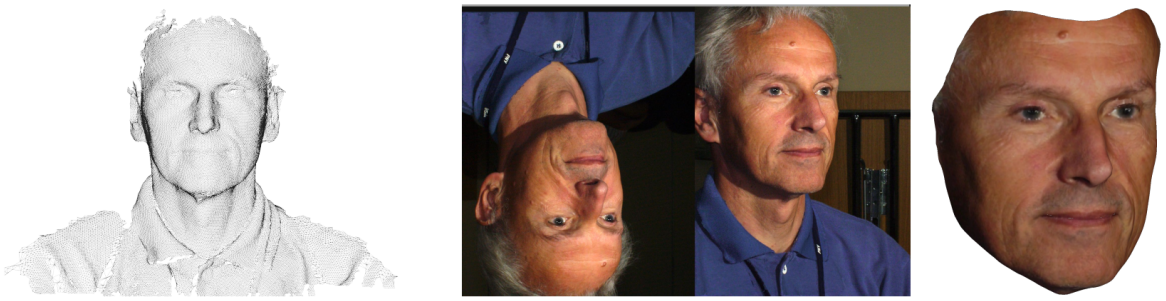

Figure 3.2: Éléments extraits du maillage brut du logiciel 3dMDface [31] : (au centre) Photographie texturée sous deux angles capturée par les caméras 3dMDface. (à droite) : Le scan et la texture sont intégrés étroitement dans le modèle 3D.

Le processus de création du M3DFM commence par la construction d'un ensemble de formes faciales à partir d'un grand nombre de scans 3D de visages. Vous pouvez voir un exemple de scan dans la Figure 3.2 qui présente le maillage capturé à gauche, la texture RVB au milieu, et le scan après recalage sur le modèle 3D. Les auteurs utilisent ensuite une technique d'analyse en composantes principales (PCA) pour réduire la dimension de l'espace de formes et obtenir un espace latent compact et contrôlable. De plus, pour capturer les variations fines de texture, le modèle utilise des cartes de texture séparées pour chaque niveau de résolution.

Booth [32] présentent une nouvelle approche pour la création de modèles déformables tridimensionnels à grande échelle, également connus sous le nom de "Large Scale 3D Morphable Models" (Large Scale 3D Morphable Models (LS3DMM)). Les modèles déformables tridimensionnels sont largement utilisés pour la modélisation et la génération de

formes tridimensionnelles, notamment pour les visages humains. Cependant, la plupart des modèles existants souffrent de limitations en termes de taille et de variabilité, ce qui limite leur application à grande échelle. Dans cet article, les auteurs proposent une méthodologie innovante pour construire des LS3DMM à grande échelle, en s'appuyant sur un vaste ensemble de données de formes faciales tridimensionnelles provenant de différentes sources. Pour construire le LS3DMM, les auteurs collectent une énorme base de données comprenant des visages 3D provenant de différentes sources et de diverses, assurant ainsi une grande diversité des caractéristiques faciales. En utilisant cette base de données, ils décrivent les étapes clés pour la construction du modèle, notamment le prétraitement des données, la réduction de dimension, l'alignement des visages et la décomposition en valeurs singulières (SVD) pour représenter les variations.

$$\mathbf{V} = \overline{\mathbf{V}} + \sum_{i=1}^{n} c_i \mathbf{b}_i^s + \sum_{i=1}^{n} d_i \mathbf{b}_i^t \tag{3.1}$$

Où :

- $\mathbf{V}$ est le modèle de visage généré en 3D.
- $\bar{\mathbf{V}}$ est la forme moyenne du visage.
- c_i et d_i sont les coefficients de déformation pour les bases de formes et de textures respectivement.
- $\mathbf{b}_i^s$ et $\mathbf{b}_i^t$ sont les vecteurs de base pour les formes et les textures.

Les expériences menées par les auteurs montrent que le LS3DMM atteint une excellente performance pour la reconnaissance faciale, l'alignement et le suivi des visages, démontrant ainsi son efficacité et son utilité dans des applications pratiques à grande échelle.

3.3 Reconstruction 3D du visage à partir d'une forme géométrique

3.3.1 Modèles antagoniste pour la reconstruction 3D du visage

La reconstruction 3D du visage est une tâche complexe et essentielle dans de nombreux domaines tels que la réalité virtuelle, l'animation, la surveillance, la reconnaissance faciale, et bien d'autres. Les modèles antagonistes, également connus sous le nom de GAN (Generative Adversarial Networks), ont révolutionné le domaine de l'apprentissage profond

en permettant la génération de données réalistes et la reconstruction précise à partir d'informations incomplètes. Dans cette section de l'état de l'art, nous allons explorer l'utilisation des modèles antagonistes pour la reconstruction 3D du visage. Dans cette section, nous examinerons les travaux de recherche récents qui ont utilisé des modèles antagonistes pour la reconstruction 3D du visage. Nous discuterons des architectures de GAN spécifiquement conçues pour cette tâche.

Deng, Y., et al [33] proposent une méthode de reconstruction de visages à partir d'une image 2D en exploitant des informations d'image hybrides pour l'apprentissage faiblement supervisé. La figure 3.3 présente l'architecture globale. Ils utilisent une combinaison de pertes à différents niveaux pour la reconstruction. Ces pertes comprennent une perte robuste au niveau de l'image et une perte au niveau de la perception. La combinaison de ces pertes permet d'obtenir des résultats plus précis que les méthodes précédentes, qui étaient entraînées de manière entièrement supervisée. La seconde contribution concerne l'agrégation des reconstructions de visages à partir de plusieurs images. Les auteurs proposent un schéma d'apprentissage de confiance pour prédire les coefficients de modèle 3D du visage avec une meilleure précision. Ils entraînent un réseau auxiliaire pour produire des "scores de confiance" des coefficients de modèle 3D du visage prédits, puis ils utilisent ces scores pour effectuer une agrégation basée sur la confiance. Cette approche permet d'améliorer la qualité des reconstructions en favorisant les photos de haute qualité et visibilité, tout en exploitant les différences de pose pour fusionner de manière plus précise les informations complémentaires provenant des différentes images.

L'approche proposée surpasse les méthodes précédentes sur plusieurs ensembles de données, avec des performances significativement meilleures, même en utilisant un espace de représentation tridimensionnelle de faible dimension.

Triple-GAN [34] est une méthode d'apprentissage profond utilisée pour réaliser la progression du vieillissement du visage. L'objectif principal est de prendre une image d'un visage à un certain âge et de la faire progresser pour représenter l'apparence du même visage à un âge plus avancé. Cela peut être utile dans divers domaines, tels que les applications de retouche photo, la création de personnages pour des films ou jeux vidéo, etc.

L'approche proposée surpasse les méthodes précédentes sur plusieurs ensembles de données, avec des performances significativement meilleures, même en utilisant un espace de représentation tridimensionnelle de faible dimension.

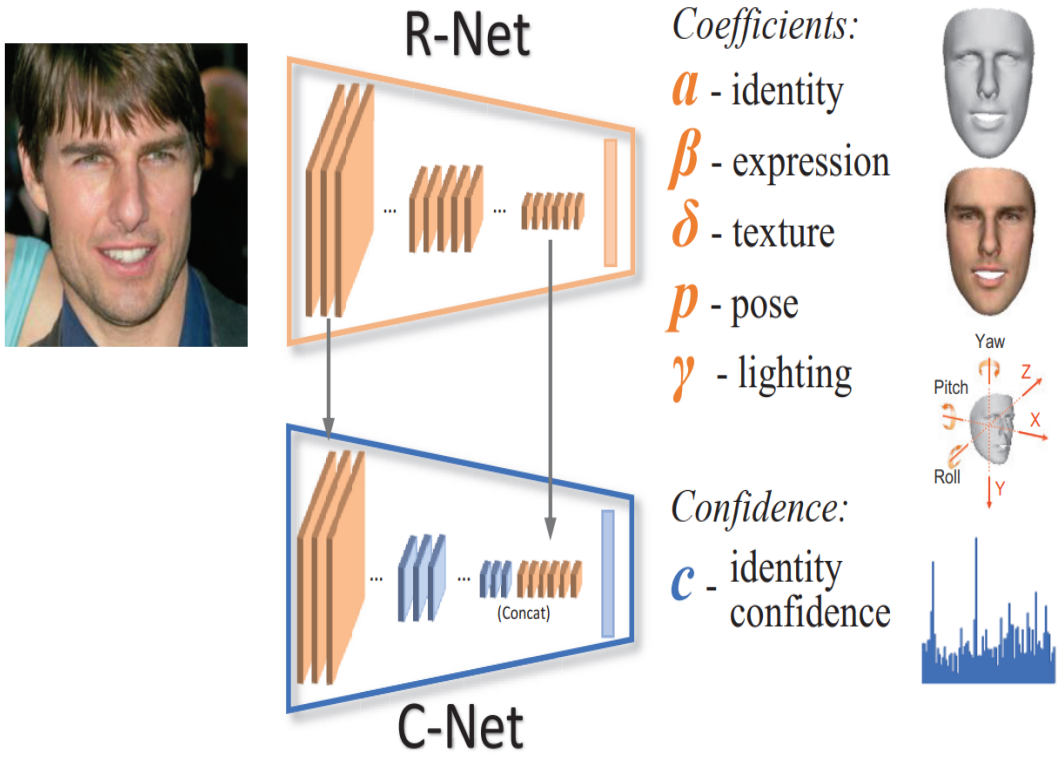

Figure 3.3: Architecture global du modèle [33]

La méthode "Triple-GAN" utilise des réseaux de neurones adversaires génératifs (GAN) pour accomplir cette tâche. Les GAN sont composés de deux réseaux de neurones en compétition : un générateur et un discriminateur. Le générateur tente de créer des images réalistes, tandis que le discriminateur essaie de les différencier des images réelles. Au fur et à mesure de l'entraînement, le générateur devient de plus en plus compétent pour produire des images convaincantes qui peuvent tromper le discriminateur.

Ce qui distingue le "Triple-GAN" des approches classiques de progression de l'âge, c'est l'utilisation d'une "perte de traduction triple" (triple translation loss) pour guider l'apprentissage du générateur. Cette perte vise à aligner les caractéristiques sémantiques du visage à différents âges. Elle utilise trois traductions différentes pour garantir une progression en douceur de l'âge tout en conservant les attributs caractéristiques du visage.

Les trois traductions sont : 1) traduction de l'image jeune à l'image âgée, 2) traduction de l'image âgée à l'image jeune, et 3) traduction de l'image jeune à l'image reconstruite et ensuite à l'image âgée. Cette approche triple contribue à la stabilité de l'apprentissage

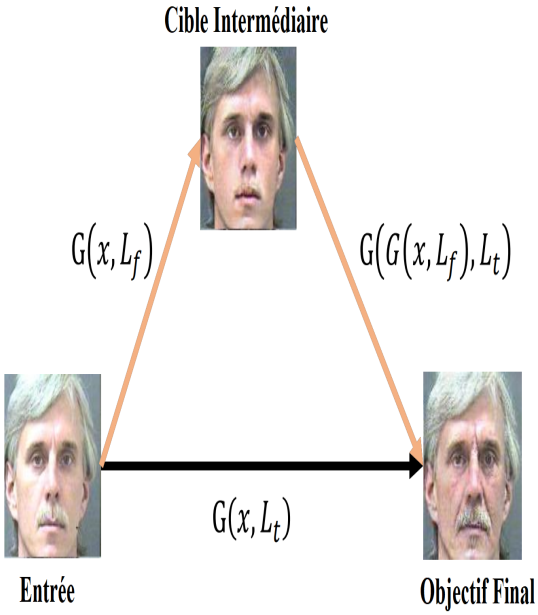

Figure 3.4: Illustration de l'utilisation de différents générateurs appliqués pour traduire les visages synthétisés d'un groupe d'âge particulier vers un autre groupe d'âge dans [34] .

et à la qualité des résultats de la progression de l'âge. Triple-GAN permet d'obtenir des résultats de haute qualité en conservant les caractéristiques uniques du visage tout en simulant l'effet du vieillissement. La figure 3.4 illustre l'utilisation de différents générateurs appliqués.

Le principe de la méthode proposée dans [35] consiste à introduire une nouvelle approche pour la reconstruction 3D à partir de réseaux adversaires génératifs (GAN) entrainés à l'aide d'images 2D non étiquetées. L'objectif principal est de déterminer si les GAN 2D peuvent capturer des informations tridimensionnelles implicites et être utilisés pour la reconstruction 3D sans supervision, c'est-à-dire sans l'utilisation de données 3D annotées.

La méthode se base sur deux étapes clés : l'entraînement d'un GAN 2D et la reconstruction 3D à partir des sorties du GAN.

Le générateur du GAN prend des vecteurs de bruit aléatoire en entrée et génère des images 2D synthétiques. Le discriminateur est formé pour distinguer entre les images réelles du jeu de données et les images générées par le générateur. La fonction de perte

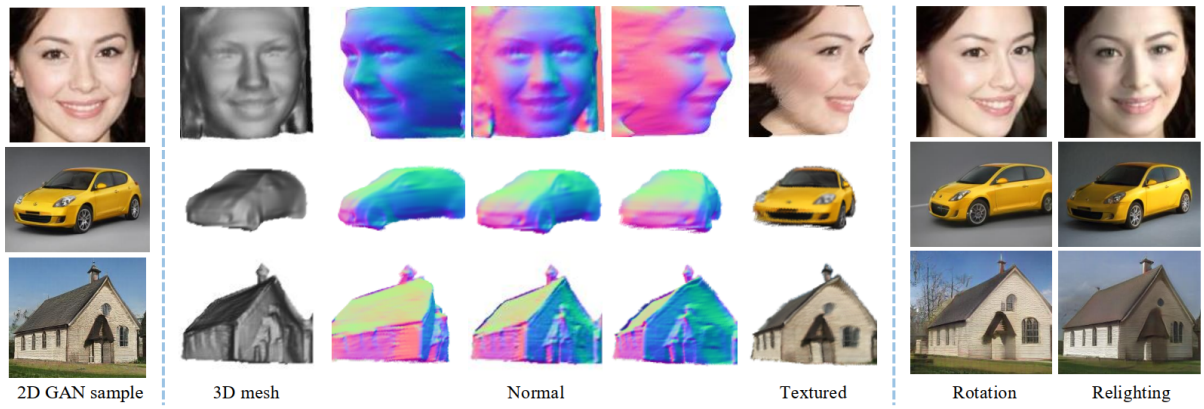

Figure 3.5: Résultats générés par le modèle [35] montre la capacité du modèle à reconstruire de manière non supervisée la forme en 3D (visualisée sous forme de maillage 3D, de normales de surface et de texture)

utilisée pour l'entraînement du GAN est une combinaison de deux termes : l'erreur de reconstruction et l'erreur d'adversité. L'erreur de reconstruction mesure la différence entre les images générées et les images réelles dans l'espace des pixels. L'erreur d'adversité vise à maximiser la capacité du générateur à tromper le discriminateur. Le GAN est formé de manière itérative jusqu'à ce qu'une convergence satisfaisante soit atteinte. Une fois le GAN entraîné, les images 2D générées par le générateur sont utilisées comme données d'entrée pour la reconstruction 3D. Pour cela, une architecture spécifique de réseau de neurones est utilisée pour effectuer la reconstruction. Le réseau prend les images 2D en entrée et prédit la représentation 3D de la forme latente correspondante. La fonction de perte utilisée pour la reconstruction 3D mesure la différence entre la représentation 3D prédite et une estimation de la véritable forme 3D, mais sans utiliser de données 3D annotées. Cette perte encourage le réseau à apprendre des caractéristiques 3D significatives à partir des images 2D sans supervision.

Les résultats expérimentaux montrent que la méthode proposée parvient à reconstruire des formes 3D réalistes à partir d'images 2D générées par le GAN 2D, sans nécessiter de données 3D étiquetées. Les expériences de validation qualitative et quantitative démontrent que les informations 3D apprises implicitement par le GAN 2D permettent une reconstruction précise des formes 3D. La Figure 3.5 présente les résultats générés par le modèle."

Zhu, X. [36] les auteurs proposent une approche complète pour l'alignement facial en 3D

sur une gamme complète de poses du visage. L'alignement facial est une tâche cruciale dans la vision par ordinateur et la reconnaissance faciale, qui consiste à localiser les points clés du visage (comme les yeux, le nez, la bouche) dans une image et à les repositionner de manière cohérente pour une analyse ultérieure. La figure 3.6 illustre les résultats générés par le modèle.

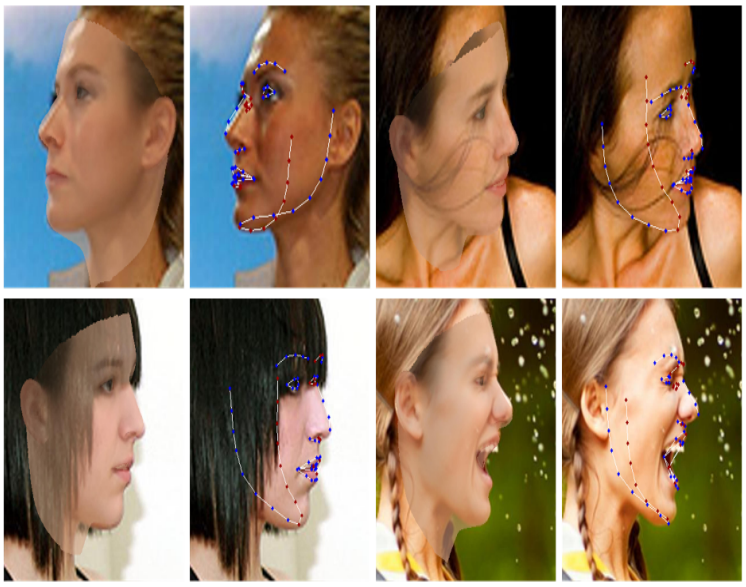

Figure 3.6: Résultats d'allignement généré par le modèle [36]

Leur méthode se compose de trois étapes principales :

- Localisation des points clés 2D du visage : Tout d'abord, un détecteur de visage est utilisé pour extraire la région du visage à partir de l'image. Ensuite, un modèle de points clés 2D est utilisé pour localiser les positions approximatives des points clés du visage.
- Estimation des paramètres 3D : Pour chaque image d'entrée, la méthode estime les paramètres 3D qui décrivent la pose globale du visage. Ces paramètres comprennent la rotation, la translation et l'échelle du visage dans l'espace 3D.
- Projection en 3D des points clés 2D : En utilisant les paramètres 3D estimés, les points clés 2D détectés sont projetés dans l'espace 3D pour obtenir leur position en 3D.

La méthode utilise un modèle 3D préalablement construit, appelé 3DMM, qui représente la variabilité de formes et de textures du visage. Ce modèle est utilisé pour estimer les paramètres 3D du visage à partir des points clés 2D détectés.

Les auteurs ont également introduit une nouvelle technique pour gérer les poses extrêmes, qui sont souvent difficiles à aligner avec les approches traditionnelles. Ils ont introduit une version augmentée du 3DMM, appelée "Shape Incremental PCA" (SI-PCA), qui permet une meilleure représentation des poses extrêmes et garantit une meilleure précision d'alignement dans toute la gamme de poses.

La méthode a été évaluée sur plusieurs ensembles de données standard et a montré des performances supérieures par rapport aux méthodes existantes, en particulier pour les poses extrêmes du visage. Elle a démontré sa robustesse et son efficacité pour l'alignement facial en 3D, ouvrant ainsi de nouvelles possibilités pour la reconnaissance faciale et d'autres applications liées à la vision par ordinateur.

Le Joint 3D Face Reconstruction and Dense Alignment with Position Map Regression (PRNet) [37] est un réseau de neurones profond spécialement conçu pour estimer simultanément la géométrie 3D du visage et l'alignement précis des points caractéristiques. La figure 3.7 présente l'architecture globale de ce modèle. Les points caractéristiques comprennent des éléments tels que les coins des yeux, le nez et la bouche. Cette approche diffère des méthodes traditionnelles qui traitent généralement la reconstruction 3D et l'alignement 2D-3D comme des étapes distinctes. Les auteurs introduisent une architecture de réseau neuronal spécifique (PMRN) qui permet de prédire simultanément la géométrie tridimensionnelle du visage et l'alignement des points caractéristiques.

Une représentation innovante sous forme de "cartes de position" est utilisée pour capturer la géométrie du visage et l'alignement des points. Cette représentation se révèle être plus précise et informative que les approches traditionnelles. La contribution de cet article met en évidence les avantages de l'apprentissage conjoint de la reconstruction 3D du visage et de l'alignement dense, ce qui améliore considérablement la précision globale du modèle. Les auteurs présentent une évaluation approfondie de leur modèle en comparaison avec d'autres méthodes de pointe. Les résultats montrent que l'approche PMRN surpasse de manière significative les techniques existantes en termes de précision de la reconstruction 3D et de l'alignement des points caractéristiques.

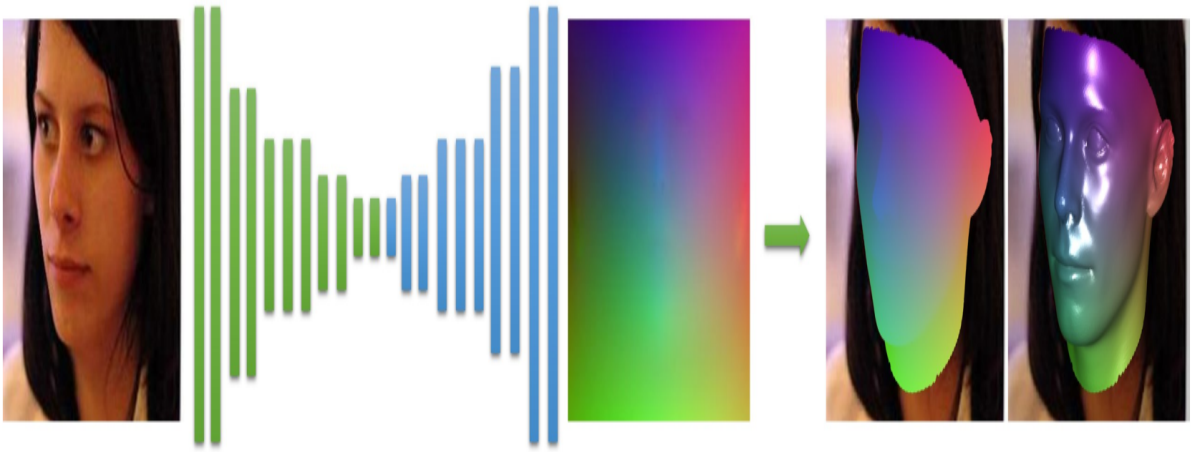

Figure 3.7: L'architecture du modèle [37]

L'objectif principal de l'article Detailed Expression Capture and Animation (DECA) [38] est de développer un modèle 3D détaillé et animable des visages humains en utilisant un ensemble de données d'images provenant de diverses sources et conditions d'éclairage, ce qui simule davantage les situations que l'on peut rencontrer dans la vie quotidienne. Les auteurs ont rassemblé un ensemble de données massif de visages à partir d'images en conditions réelles, notamment des photos de personnes dans des environnements variés. Ce vaste ensemble de données a servi de base à leur modèle. Les auteurs ont conçu un modèle génératif profond qui est capable de créer des modèles 3D de visages humains à partir d'une image en 2D. Le modèle est conçu pour être animable, ce qui signifie qu'il peut générer des visages dans diverses expressions.
L'apprentissage non supervisé a été utilisé pour entraîner le modèle à partir des données.

Cela signifie que le modèle a appris à extraire des caractéristiques et à générer des visages 3D sans avoir besoin d'étiquettes de données spécifiques. Les auteurs ont évalué leur modèle en le comparant à d'autres méthodes de modélisation 3D des visages. Ils ont montré que leur modèle surpasse les méthodes existantes en termes de détail et de fidélité de la représentation.

3.4 Conclusion

Ce chapitre a été consacré à l'examen approfondi de l'état de l'art en matière de reconstruction 3D, en mettant l'accent sur la reconstruction à partir d'images et la reconstruction tridimensionnelle du visage. Cette exploration des travaux précédents a permis de mettre en lumière les avancées, les méthodes et les défis qui ont jalonné ce domaine passionnant. Dans la première section, nous avons abordé la reconstruction 3D à partir d'images, en mettant en évidence deux approches majeures : les modèles déformables tridimensionnels et la Structure acquise à partir du mouvement. Les modèles déformables tridimensionnels ont été décrits comme des outils puissants pour la reconstruction 3D en utilisant des modèles paramétriques pour décrire la forme et l'apparence. De plus, la SfM, qui repose sur la géométrie épipolaire, a été soulignée comme une méthode essentielle pour reconstruire des scènes 3D à partir de multiples images 2D.

Dans la deuxième section, nous avons exploré en détail la reconstruction tridimensionnelle du visage à partir d'une forme géométrique. Nous avons notamment mis en avant les modèles antagonistes (GAN) en tant qu'approche prometteuse pour la reconstruction 3D du visage. Ces modèles ont la capacité de générer des géométries réalistes et précises du visage à partir d'images 2D, grâce à un générateur et un discriminateur qui s'entraînent mutuellement.

L'état de l'art en matière de reconstruction 3D nous a permis de constater les avancées significatives réalisées dans ce domaine, tout en mettant en évidence certaines lacunes et défis persistants. La reconstruction 3D est une discipline en constante évolution, avec des applications potentiellement révolutionnaires dans de nombreux domaines, de la réalité virtuelle à la médecine.

Le chapitre suivant, constituera la pierre angulaire de notre travail, où nous présenterons en détail notre approche pour la reconstruction 3D du visage en utilisant les modèles

génératifs GAN. Nous élaborerons sur les choix méthodologiques, les outils et les techniques que nous avons sélectionnés pour résoudre les défis identifiés dans cet état de l'art. Ce chapitre sera le pivot de notre contribution à la recherche en reconstruction 3D, offrant une perspective prometteuse pour l'avenir de ce domaine en constante évolution.

4

L'approche proposée

4.1 Introduction

La présente thèse s'inscrit dans le cadre de l'étude d'un domaine essentiel de l'intelligence artificielle : la reconstruction 3D des visages en utilisant les modèles génératifs. L'objectif principal de notre recherche est de développer un modèle de deep learning performant et robuste, capable de recostruire des visages avec une précision élevée et une efficacité optimale. Pour atteindre cet objectif ambitieux, la méthodologie adoptée se décompose en plusieurs étapes clés, chacune étant soigneusement conçue pour répondre à des aspects spécifiques de notre problématique.

Dans un premier temps, nous présenterons l'architecture de notre modèle. Il s'agit d'un réseau de neurones profonds antagoniste basé sur des couches de convolution graphique, spécialement conçu pour la reconstruction des visages 3D. Nous détaillerons les différentes couches du notre modèle, la fonction de coût spécialement conçue pour minimiser l'erreur de prédiction en termes de forme et de caractéristiques du visage. Nous décrirons en détail cette fonction de coût et ses composants afin d'optimiser les performances de notre modèle.

4.2 L'architecture du modèle

L'architecture du modèle fait référence à la structure globale et à l'organisation interne du réseau neuronal utilisé. Cette structure dicte comment les données sont traitées et comment les relations complexes entre les variables sont apprises. Dans cette section, nous allons explorer en détail l'architecture de notre modèle. Nous aborderons en détail les différents blocks composants le générateur et le discriminateur, les, les couches du modèle, et les hyperparamètres utilisés pour optimiser le modèle.

4.2.1 Générateur

Le générateur de notre modèle est basé sur une approche existante décrite dans [39]. Cette approche prend en entrée une seule image 2D ainsi que des points de repère et produit en sortie un modèle géométrique 3D du visage correspondant. La figure 4.1 illustre notre générateur en détail. Nous avons adapté cette approche en construisant notre propre générateur qui utilise des blocs de reconstruction (reconstruction blocks) suivis de couches de triangulation de maillage (triangulation mesh layers), sauf pour la dernière couche. Plus précisément, notre générateur est composé de quatre blocs de reconstruction, chacun suivi d'une couche de triangulation de maillage. Cette conception permet de construire

progressivement un modèle 3D plus précis à chaque bloc, en utilisant les informations de l'image 2D et des points de repère.

Pour l'extraction des caractéristiques de l'image d'entrée, nous avons utilisé les quatre premiers blocs du réseau de neurones ResNet-50 comme backbone. Le réseau ResNet-50 est largement utilisé dans plusieurs travaux et a montré de bonnes performances pour la classification d'images. En combinant les blocs de reconstruction, les couches de triangulation de maillage et le backbone ResNet-50, notre générateur peut produire un modèle 3D précis et détaillé du visage à partir d'une seule image 2D et des points de repère correspondants. L'architecture de notre générateur est présentée dans la figure ci-dessous.

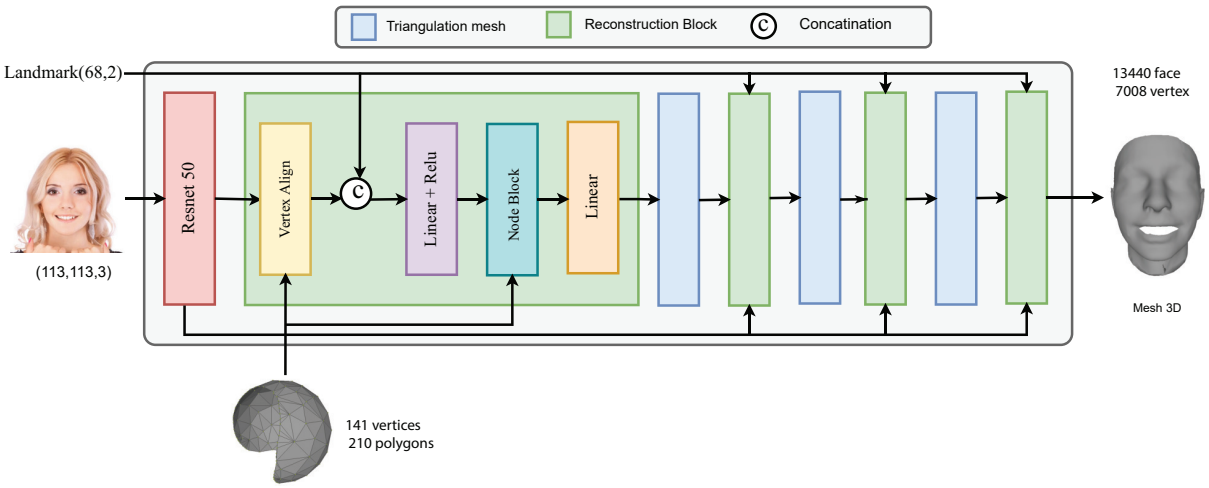

Figure 4.1: Architecture de notre générateur proposé

Avec cette architecture, nous cherchons à améliorer la qualité et la fidélité de la reconstruction 3D des visages à partir d'une seule image 2D, tout en minimisant la complexité et les ressources nécessaires pour l'entraînement et l'inférence.

4.2.1.1 Reconstruction Block

Le bloc de reconstruction de notre générateur joue un role cruciale dans le processus de génération de la géométrie du visage. Ce bloc est composé de deux opérations majeures : le vertex alignement et le NODE Block. Le vertex alignement vise à établir une correspondance entre les caractéristiques de l'image et le modèle de visage 3D. Pour ce faire, nous nous inspirons des travaux de [39], [40] et appliquons une méthode qui consiste à aligner les sommets du modèle 3D avec les caractéristiques de l'image en effectuant une

correspondance point à point.

Le modèle 3D utilisé dans le processus de vertex alignement est un semi-sphère composé de 141 sommets et 210 polygones. L'objectif principal de ce processus d'alignement est de trouver la transformation qui permet de projeter les sommets du modèle 3D sur les caractéristiques de l'image 2D, de manière à obtenir une adéquation optimale entre le modèle et les données de l'image.

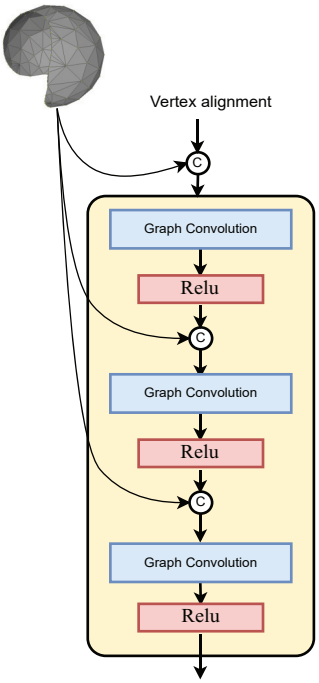

Figure 4.2: Architecture du blocs de nœuds.

Une fois le vertex alignement effectué, sa sortie est concaténée avec les points de repère du visage qui ont été extraits précédemment. Cette concaténation est ensuite utilisée en tant qu'entrée pour une couche linéaire, suivie d'une fonction d'activation ReLU. Cette étape vise à fusionner les informations extraites du modèle 3D et les caractéristiques du visage provenant de l'image 2D.

En effet la section de reconstruction block de notre générateur combine le processus de vertex alignement pour correspondre les caractéristiques 3D du modèle de visage avec les données 2D de l'image, puis concatène ces informations avec les points de repère du visage pour créer une représentation complète et enrichie du visage. Cette représentation est ensuite utilisée pour générer un visage réaliste et cohérent.

Node Block

Le Node-block est une étape essentielle dans le traitement du graphe pour agréger les caractéristiques de l'image ainsi que les points de repère du visage après l'alignement des sommets. Cette opération repose sur l'utilisation d'un réseau de neurones spécialisé dans le traitement de données structurées. Le Node-block est constitué de trois couches de convolution graphique, chacune étant suivie d'une fonction d'activation ReLU, comme illustré dans la figure 4.2. Pour la première couche de convolution graphique, l'entrée est la sortie de la couche linéaire qui résulte de l'alignement des sommets et qui est passée à travers une fonction d'activation ReLU. Pour les couches suivantes, l'entrée est constituée de la concaténation entre la sortie de la couche précédente et le modèle de visage 3D utilisé lors de l'alignement des sommets.

La sortie de chaque couche de convolution graphique est ensuite transmise à la couche suivante en suivant le même processus. Ces couches de convolution graphique permettent d'agréger les informations provenant des différentes parties du visage tout en prenant en compte leur structure et leur relation les unes avec les autres. Ensuite, ces caractéristiques agrégées passent par une couche linéaire dont l'objectif est de les transformer en une représentation 3D du visage.

Le Node-block joue un rôle crucial en capturant les relations spatiales entre les différentes parties du visage, ce qui est essentiel pour reconstruire en 3D un visage à partir d'une simple image 2D.

4.2.1.2 Triangulation

La couche de triangulation de notre générateur joue un rôle essentiel dans la construction de notre modèle de reconstruction 3D des visages. Son objectif est de transformer un maillage de départ relativement simple en un maillage plus complexe, comprenant davantage de faces, de sommets et d'arêtes. Voici comment cette couche de triangulation fonctionne :

- Insertion de sommets : Le processus de triangulation commence par l'insertion d'un sommet supplémentaire dans le milieu de chaque arête formant un triangle. En ajoutant ces nouveaux sommets, chaque triangle initial est divisé en quatre triangles plus petits. Cette étape est cruciale pour augmenter la densité du maillage, ce qui permettra de représenter des détails plus fins du visage.

- Connexion des triangles : Après l'ajout des nouveaux sommets, des arêtes sont tracées pour connecter ces sommets entre eux (voir la figure 4.3), formant ainsi les nouveaux triangles. Cette étape permet de créer des liaisons entre les triangles nouvellement formés, assurant ainsi la cohérence et l'intégrité du maillage.
- Raffinement du maillage : À la fin du processus de triangulation, nous obtenons un maillage plus complexe, avec des détails plus fins et des bords plus réguliers. Ce raffinement du maillage est essentiel pour représenter de manière plus précise les contours et les formes du visage, ainsi que les détails tels que les plis et les ridules de la peau.
- Positionnement dans le modèle : Les couches de triangulation sont ajoutées à notre modèle après la reconstruction block. Cela signifie que la triangulation intervient à un stade avancé de la reconstruction 3D, une fois que le modèle a déjà acquis certaines caractéristiques basiques du visage à partir de l'image 2D d'entrée.

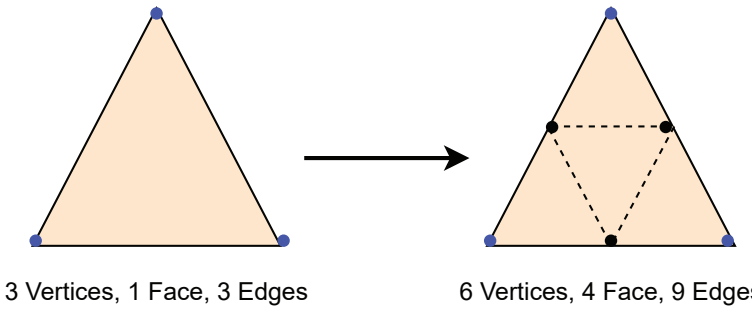

Figure 4.3: La transformation Mesh après une étape de triangulation.

En combinant la triangulation avec les autres blocs de reconstruction, notre modèle devient capable de produire des modèles 3D de visages qui sont à la fois précis, détaillés et fidèles à l'image d'entrée. La triangulation permet d'obtenir un maillage 3D plus détaillé, améliorant ainsi la qualité et la résolution des modèles finaux.

4.2.2 Le discriminateur

Le discriminateur est un élément essentiel dans notre approche de deep learning visant à la reconstruction 3D des visages. Son rôle principal est de différencier les modèles générés par notre générateur de ceux présents dans le jeu de données d'apprentissage. Pour ce faire, le discriminateur est mis en place sous la forme d'un réseau de neurones, qui suit une architecture spécifique.

Le réseau de neurones du discriminateur est constitué de deux parties principales : une couche de graphe convolution simple et un block GCN. La première partie, la couche de graphe convolution simple, est responsable d'effectuer une première transformation des caractéristiques d'entrée. Ensuite, cette sortie est transmise au block GCN.
Le block GCN est composé de cinq couches de GCN. Chaque couche est connectée à l'entrée et à la sortie du couche précédente via une opération de concaténation, comme illustré dans la figure 4.4. Cela permet d'incorporer des informations de différentes profondeurs dans le réseau, ce qui peut être crucial pour la prise de décision du discriminateur. La sortie de chaque couche GCN est de taille 16, et la taille finale de la sortie du block GCN est de 96.

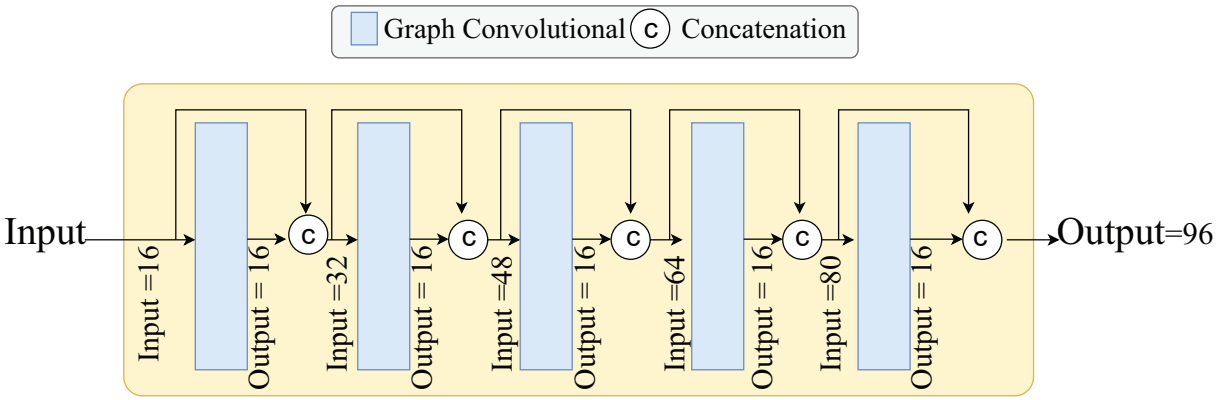

Figure 4.4: L'architecture plus détaillée du bloc GCN.

Après l'étape du block GCN, les caractéristiques sont extraites. Pour obtenir une valeur unique représentative des caractéristiques, une fonction de "mean node" est appliquée. Cette fonction calcule la valeur moyenne des caractéristiques extraites par le block GCN.

Enfin, pour obtenir une sortie binaire, indiquant si le modèle d'entrée est réel ou faux, une couche linéaire est appliquée après la fonction de "mean node". La figure 4.5 présente l'architecture globale de notre discriminateur."

Le discriminateur est entraîné avec deux types d'entrées : soit un échantillon généré par le générateur, soit un échantillon réel provenant du jeu de données d'apprentissage. L'objectif du processus d'entraînement est de minimiser la différence entre les sorties produites par le discriminateur pour les échantillons générés et réels. Cette minimisation vise à améliorer la qualité de la reconstruction réalisée par le générateur. En d'autres termes, le discriminateur

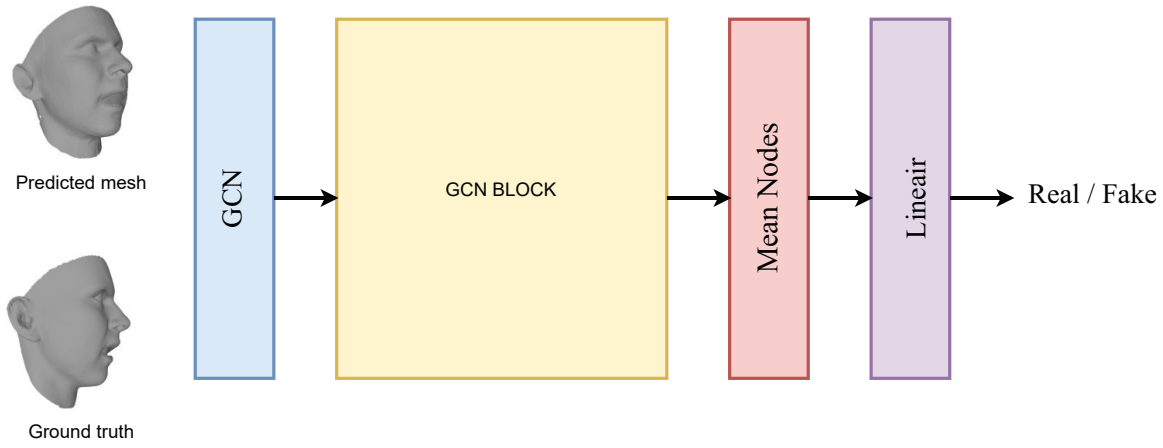

Figure 4.5: L'architecture de notre discriminateur.

apprend à mieux distinguer les vraies données des fausses données générées par le générateur, ce qui permet à ce dernier de s'améliorer dans la reconstruction 3D des visages.

4.3 La fonction de coût

La fonction de coût utilisée dans notre approche pour la reconstruction 3D des visages est constituée de plusieurs termes. Chacun de ces termes contribue à guider le processus d'apprentissage de notre modèle afin de produire des résultats de reconstruction de meilleure qualité. Les différents termes de la fonction de coût que nous utilisons sont : la perte d'adversité générale, la fonction de distance de Chamfer, la perte de distance normale, la régularisation laplacienne et la régularisation de la longueur des bords (Edge Loss). Tous ces termes de coût sont combinés de manière pondérée pour former une fonction de coût globale. Cette fonction de coût globale est alors utilisée pour guider l'optimisation de notre réseau de neurones lors de la phase d'apprentissage. En ajustant les poids de ces différents termes, nous pouvons influencer le comportement de notre modèle et améliorer la qualité de la reconstruction 3D des visages.

4.3.1 La distance de Chamfer

La distance de Chamfer est une mesure couramment utilisée pour évaluer la similarité entre deux ensembles de points dans l'espace. Dans le contexte de la reconstruction 3D des visages à l'aide de GANs, la distance de Chamfer est utilisée pour comparer le modèle

généré (sortie du GAN) avec le modèle réel (ensemble de points représentant le visage réel).

Supposons que nous avons deux ensembles de points dans un espace tridimensionnel:

- L'ensemble de points réels représentant un visage réel : $R = \{(x_i, y_i, z_i)\}$ où $i = 1, 2, \ldots, n$.
- L'ensemble de points générés par le GAN représentant le visage reconstruit : $G = \{(u_j, v_j, w_j)\}$ où $j = 1, 2, \ldots, \mathrm{m}$.

La distance de Chamfer entre ces deux ensembles de points, notée $D(R, G)$, peut être calculée comme suit :

$$\mathrm{D(R, G)} = (1/\mathrm{n}) * \Sigma_{\mathrm{i}} \min_{\mathrm{k}} ||\mathrm{R_i} - \mathrm{G_k}|| + (1/\mathrm{m}) * \Sigma_{\mathrm{j}} \min_{\mathrm{1}} ||G_{\mathrm{j}} - \mathrm{R_1}|| \tag{4.1}$$

Où $||.||$ représente la norme euclidienne (distance euclidienne) entre deux points, $\min_{\mathrm{k}}$ indique le point le plus proche de R_i dans G, et min_l indique le point le plus proche de G_j dans R. La distance de Chamfer calcule donc la moyenne des distances entre chaque point de l'ensemble réel et son point le plus proche dans l'ensemble généré, ainsi que la moyenne des distances entre chaque point de l'ensemble généré et son point le plus proche dans l'ensemble réel.

L'objectif lors de la reconstruction 3D des visages à l'aide de GANs est de minimiser la distance de Chamfer entre l'ensemble réel et l'ensemble généré. Une distance de Chamfer plus petite indique une meilleure similarité entre les ensembles de points et, par conséquent, une meilleure reconstruction du visage.

Il est important de noter que le calcul de la distance de Chamfer peut être coûteux en termes de temps de calcul, car il nécessite de comparer chaque point de l'ensemble réel avec tous les points de l'ensemble généré et vice versa. Cependant, c'est une mesure utile pour évaluer la qualité de la reconstruction 3D d'un visage à l'aide de GANs et permet d'optimiser le modèle pour obtenir des résultats plus réalistes.

4.3.2 La normale distance

La normale distance est un concept essentiel en reconstruction 3D des visages, notamment dans le domaine de la vision par ordinateur et de la modélisation 3D. Elle permet d'estimer la distance d'un point donné d'une surface à un plan tangent, ce qui est particulièrement

utile pour créer des modèles tridimensionnels réalistes des visages.

Dans le contexte de la reconstruction 3D des visages, son principe repose sur la comparaison des directions angulaires des normales entre les surfaces générées par notre modèle et celles des surfaces de référence.

Pour calculer la normale distance, nous avons besoin de la description mathématique de la surface du visage. Dans de nombreux cas, cette surface est représentée sous forme de maillage triangulé, où chaque point du maillage est défini par ses coordonnées (x, y, z) dans l'espace 3D.

$$l_{\mathrm{ND}}(P, G) = -|P|^{-1} \sum_{(p,g)\in\Lambda_{P,G}} |u_p \cdot u_g| - |G|^{-1} \sum_{(g,p)\in\Lambda_{G,P}} |u_g \cdot u_p| \tag{4.2}$$

Où $\Lambda_{P,G} = \{(p, \arg\min_g \|p - g\|) : p \in P\}$ est l'ensemble des paires (p, g) telles que $g \in G$ est le plus proche voisin de $p \in P$.

La normale distance est utilisée dans différents contextes liés à la reconstruction 3D des visages, notamment dans les techniques de numérisation 3D, la modélisation par photogrammétrie, et la réalisation de scans 3D à partir de caméras ou capteurs spécialisés.

En pratique, les points 3D du visage sont souvent obtenus à partir de différentes sources, telles que des scanners 3D, des systèmes de vision stéréoscopique ou des techniques de photogrammétrie. En utilisant les points ainsi obtenus, les normales à la surface peuvent être calculées localement pour chaque point du maillage.

En combinant les normales et les distances associées, les algorithmes de reconstruction 3D sont capables de créer des modèles tridimensionnels détaillés des visages, permettant ainsi une multitude d'applications, comme l'animation, l'analyse des expressions faciales, l'identification biométrique, ou même l'adaptation de modèles virtuels à des visages réels.

4.3.3 lissage laplacien

Une méthode couramment utilisée pour améliorer la qualité de la reconstruction consiste à appliquer le lissage Laplacien (Laplacien smoothing). Cette technique permet de régulariser la géométrie de la surface 3D en réduisant les variations abruptes de profondeur entre les points adjacents, tout en préservant les caractéristiques essentielles du visage.

Le Laplacien Smoothing est une opération mathématique appliquée à la surface 3D d'un objet, dans notre cas, le visage reconstruit. Il agit en redistribuant les coordonnées 3D des points de manière à réduire les différences de profondeur entre les points voisins. Cela a pour effet de lisser la surface tout en conservant les détails globaux du visage.

Soit S une surface 3D représentée par un ensemble de points $P = \{p_1, p_2, \ldots, p_n\}$,Pour calculer cette perte, nous définissons d'abord une coordonnée laplacienne pour chaque sommet p comme suit:

$$\delta_p = p - \sum_{i \in \mathcal{N}(p)} \frac{1}{\|\mathcal{N}(p)\|} i, \tag{4.3}$$

La régularisation laplacienne est alors calculée par la formule ci-dessous :

$$l_{smoth} = \sum_{p} \left\| \delta'_p - \delta_p \right\|_2^2, \tag{4.4}$$

où p est le sommet prédit et i est le voisin de p. $\mathcal{N}(p)$ représente l'ensemble des i sommets voisins. $\|.\|$ indique le nombre d'éléments.δ_v désigne la coordonnée laplacienne du sommet p, et δ'_p fait référence à celles mises à jour.$\|.\|_2^2$ correspond à la norme L2.

En appliquant itérativement un lissage laplacien aux sommes de maillage 3D, la surface du visage est progressivement régularisée, éliminant les petites imperfections tout en préservant les détails importants. .

4.3.4 Perte des contours

Nous définissons une régularisation de la longueur des arêtes pondérées afin de contrôler la densité des sommets, car les régions centrales d'une reconstruction de visage humain nécessitent plus de sommets pour être reconstruites que les autres zones. Cela impose une arête plus courte pour une région avec un poids plus élevé, ce qui revient à rassembler plus

de sommets localement pour obtenir de meilleurs détails. La régularisation de la longueur des arêtes permet une reconstruction de qualité des arêtes composant l'objet généré. La formule suivante calcule sa valeur :

$$l_{edge} = \sum_{p} \sum_{i \in \mathcal{N}(p)} \|p - i\|_2^2 \tag{4.5}$$

Cette régularisation vise à garantir une distribution optimale des sommets pour la reconstruction, en favorisant la précision dans les régions cruciales du visage humain. Elle contribue ainsi à une reconstruction fidèle des bords qui composent l'objet généré.

4.3.5 Erreur Globale

La valeur finale de notre fonction objective est la combinaison de la fonction de cout du réseau de neurones génératif, Distance de chamfer, Normale Distance, Laplacien smooth et Edge Loss.

$$\mathcal{L}_{\text{Recon}}(G) = \mathcal{L}_{CD} + \lambda_1 \mathcal{L}_{ND} + \lambda_2 \mathcal{L}_{\text{smoth}} + \lambda_3 \mathcal{L}_{\text{edge}} \tag{4.6}$$

$$G^* = \arg\min_{G} \max_{D} \lambda \mathcal{L}_{GAN}(G, D) + \mathcal{L}_{\text{Recon}}(G) \tag{4.7}$$

4.4 Conclusion

Dans ce chapitre, nous avons plongé dans les détails de l'architecture de notre modèle, qui est au cœur de notre approche de reconstruction 3D. Le générateur et le discriminateur, deux composants essentiels de notre modèle, ont été présentés en détail. Le générateur est chargé de créer une représentation tridimensionnelle à partir d'une seule image 2D et de ses repères faciaux, tandis que le discriminateur évalue la qualité de cette représentation.

La méthodologie a également inclus une discussion approfondie sur la fonction de coût que nous utilisons pour entraîner notre modèle. Cette fonction de coût combine plusieurs composantes, dont la distance de Chamfer, la normale distance, le Laplacien smoothing et l'Edge Loss. Chacune de ces composantes joue un rôle clé dans la formation de notre modèle et contribue à l'obtention de résultats de haute qualité.

Le prochain chapitre de notre thèse sera dédié à la présentation et à l'analyse des résultats de notre approche de Reconstruction 3D. Nous examinerons en détail la performance de notre modèle par le biais de mesures quantitatives et qualitatives. Les résultats seront présentés de manière à mettre en évidence les avantages et les contributions de notre approche par rapport aux approches existantes.

5

Évaluation et Expérimentations

5.1 Introduction

Dans ce chapitre nous présentons en détail l'implémentation de notre approche pour la reconstruction tridimensionnelle des visages à partir d'images en deux dimensions en utilisant des modèles génératifs. Après avoir posé les bases théoriques dans les chapitres précédents, nous entrons maintenant dans le cœur de notre recherche en décrivant en profondeur les étapes cruciales de l'implémentation, ainsi que les résultats obtenus lors de l'évaluation de notre modèle. Cette section est cruciale pour comprendre la faisabilité et l'efficacité de notre approche.

Le prétraitement des données est une étape cruciale dans tout projet d'apprentissage automatique. Nous détaillons les méthodes et les techniques que nous avons utilisées pour préparer nos données d'entraînement. Cela inclut la réduction de la taille des images d'entrée et la réduction du maillage 3D. Nous mettons en évidence l'importance de la qualité des données pour la performance de notre modèle. Nous expliquons en détail la configuration de notre modèle et les paramètres d'entraînement spécifiques que nous avons utilisés. Cette section offre un aperçu complet du processus d'apprentissage profond qui sous-tend notre approche.

Pour évaluer la performance de notre modèle, nous le comparons rigoureusement avec les méthodes de l'état de l'art en reconstruction 3D des visages. Dans cette section, nous présentons les métriques d'évaluation que nous avons utilisées pour mesurer la qualité de nos résultats. Nous réalisons une comparaison quantitative approfondie, en mettant en lumière les avantages et les inconvénients de notre approche par rapport aux méthodes existantes. De plus, nous appuyons notre évaluation par une analyse qualitative pour illustrer la qualité de nos reconstructions. Une étude d'ablation est menée pour identifier les composants clés de notre modèle qui contribuent le plus à sa performance. Nous analysons comment la suppression ou la modification de certaines parties du modèle affecte la qualité de la reconstruction. Cette analyse nous permet de mieux comprendre la robustesse de notre approche.

5.2 Prétraitement des données

Le prétraitement des données joue un rôle fondamental dans la réussite de toute tâche d'apprentissage profond, et la reconstruction 3D des visages à l'aide de modèles génératifs

GAN ne fait pas exception. Cette section détaille les étapes cruciales de prétraitement des données que nous avons entreprises avant de procéder à l'entraînement de notre modèle. Ces étapes visaient à optimiser l'efficacité du modèle tout en préservant la qualité et les détails essentiels pour la génération de maillages 3D réalistes.

5.2.1 Réduction de la Taille des Images d'Entrée

L'une des premières étapes du prétraitement des données consiste à réduire la taille des images AFLW2000-3D [36] (voir l'image 5.1 présentant un échantillon) d'entrée du générateur. Les images originales pouvaient présenter une résolution élevée, ce qui aurait augmenté considérablement le temps de calcul nécessaire à l'entraînement du modèle. Ainsi, nous avons redimensionné les images d'entrée à une taille de 113x113 pixels. Cette décision s'est avérée cruciale pour plusieurs raisons.

Premièrement, la réduction de taille permet une diminution significative des besoins en puissance de calcul, ce qui rend l'entraînement du modèle plus rapide et plus efficace. En effet, les modèles GAN, en particulier lorsqu'ils sont profonds, nécessitent un grand nombre d'opérations de reconstruction block, Triangulation mesh et d'optimisation, ce qui peut s'avérer extrêmement gourmand en ressources. La réduction de la résolution des images d'entrée a considérablement allégé cette charge computationnelle.

Deuxièmement, la réduction de taille était effectuée avec soin afin de garantir que la qualité et les détails essentiels pour la génération de maillages 3D réalistes ne soient pas compromis. Contrairement à une simple réduction de résolution qui pourrait entraîner une perte de détails, notre processus de redimensionnement visait à préserver autant que possible les caractéristiques importantes des visages, notamment les traits distinctifs et les expressions faciales. Cette balance entre la réduction de taille et la préservation de la qualité était essentielle pour assurer la pertinence de notre modèle dans des applications réelles.

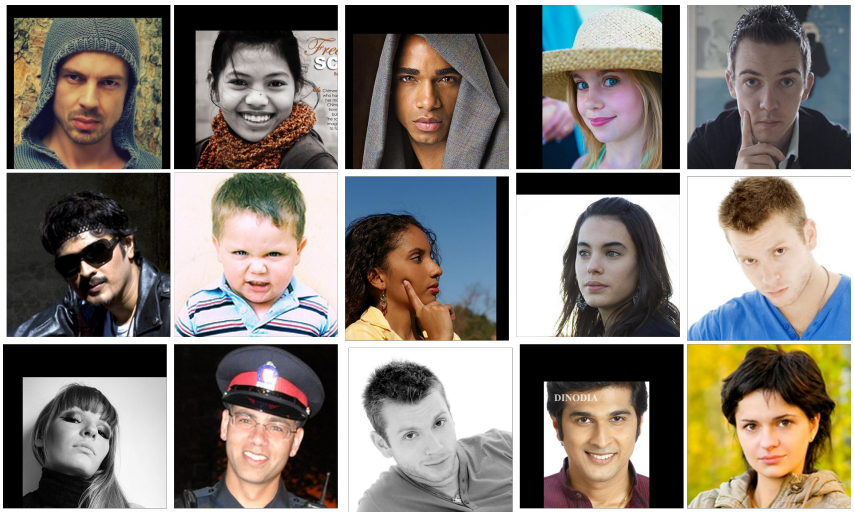

Figure 5.1: Un échantillon représentant quelques images de l'ensemble de données AFLW2000-3D utilisé pour notre modèle

5.2.2 Réduction du Maillage 3D

Outre la réduction de la taille des images d'entrée, nous avons également appliqué une réduction du maillage 3D initial utilisé dans notre modèle. Le maillage 3D initial comportait 53215 sommets et 105840 triangulations, ce qui pouvait entraîner des temps de calcul prohibitifs lors de la génération de maillages 3D complexes. Par conséquent, une réduction du maillage à 6910 sommets et 13440 triangulations a été entreprise. La figure 5.2 illustre l'objet 3D avant et après cette réduction.

La réduction du maillage 3D visait à atteindre plusieurs objectifs. Tout d'abord, elle a permis de réduire considérablement la complexité du modèle, ce qui s'est traduit par des temps de calcul plus courts et une meilleure efficacité globale. Comme pour la réduction de taille des images, cette simplification était cruciale pour rendre notre modèle plus accessible et plus performant dans des environnements avec des ressources limitées.

De plus, tout en réduisant le nombre de sommets du maillage, nous avons veillé à préserver les détails et la qualité de la géométrie 3D. Cette étape délicate a été réalisée en sélectionnant judicieusement les sommets à conserver pour assurer la fidélité de la reconstruction 3D. Le processus de réduction du maillage a ainsi été exécuté avec soin pour maintenir la précision et le réalisme de nos générations de maillages 3D.

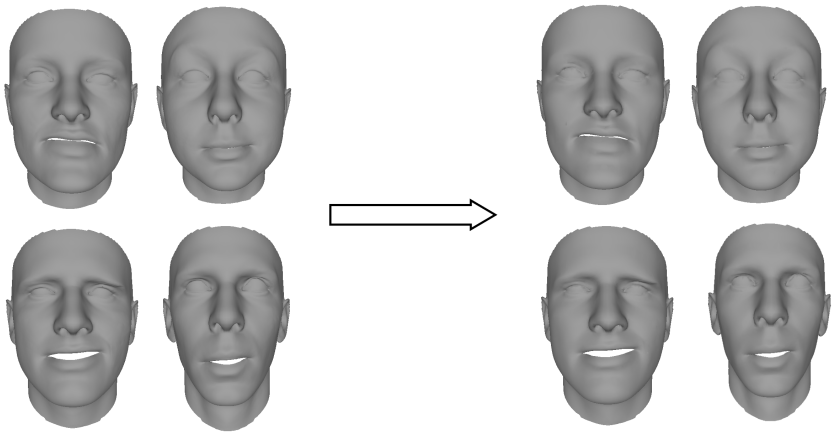

Figure 5.2: l'illustration d'un objet 3D (a) avant la réduction (b) après la réduction

5.3 Comparaison avec l'état de l'art

Dans cette section, nous procédons à une évaluation détaillée de notre modèle en effectuant une comparaison qualitative et quantitative avec les approches de l'état de l'art pour la prédiction de visage 3D à partir d'une seule image. Nous avons mené cette étude en utilisant deux ensembles de données différents : Florence [41] et AFLW2000-3D [36]. Notre objectif est de déterminer l'efficacité de notre modèle par rapport aux approches existantes : Deep3D [33], PRNet [37], DECA [38], et 3DDFA [36]. Enfin, nous avons démontré l'efficacité de chaque composant de notre architecture en réalisant une étude d'ablation.

5.3.1 Métriques de l'évaluation

Nous avons utilisé deux métriques pour évaluer les performances de notre modèle : la Distance de Chamfer (Chamfer Distance (CD)) et la Earth Mover's Distance (Earth Mover's Distance (EMD)) [42].

5.3.1.1 La Distance de Chamfer

Une valeur de Distance de Chamfer plus faible indique une meilleure correspondance entre les points du nuage de points prédits et réels, ce qui signifie que le modèle est capable de reproduire fidèlement la géométrie 3D du visage. En revanche, une valeur plus élevée de Distance de Chamfer indique des divergences significatives entre les points prédits et les points réels, ce qui témoigne d'une moindre précision du modèle dans la prédiction des détails du visage en 3D.

Lors de l'évaluation d'un modèle de prédiction de visage 3D, la Distance de Chamfer est souvent utilisée en conjonction avec d'autres métriques pour obtenir une évaluation globale de la performance du modèle. En particulier, elle est souvent associée à la Earth Mover's Distance (EMD) pour fournir une vue complète de la qualité des prédictions du modèle.

5.3.1.2 Earth Mover's Distance (EMD)

L'Earth Mover's Distance (EMD), est une mesure de dissimilarité qui trouve des applications importantes dans le domaine de l'analyse de données en trois dimensions (3D). Elle permet de quantifier la différence entre deux distributions de points tridimensionnels. Contrairement à d'autres métriques de distance, l'EMD prend en considération la géométrie spatiale ainsi que les relations de voisinage entre les points, ce qui en fait un outil puissant pour comparer des ensembles de données complexes tels que les nuages de points 3D.

L'EMD se révèle particulièrement utile pour des tâches telles que l'appariement de modèles 3D, la segmentation de nuages de points et la reconnaissance d'objets dans le contexte des données en trois dimensions. Par exemple, l'EMD peut aider à évaluer la similitude entre différentes prises de vue ou entre des modèles numériques reconstruits à partir de sources diverses lors de la comparaison de nuages de points 3D issus de scans de surfaces. Dans des domaines tels que la vision par ordinateur, la robotique et la modélisation environnementale, il est également utilisé pour calculer les différences géométriques entre des objets tridimensionnels.

L'Earth Mover's Distance (EMD) est une métrique de similarité entre deux ensembles de points, souvent utilisée dans le contexte de la complétion de nuage de points. Pour deux nuages de points S_1 et S_2, l'EMD est définie comme suit:

$$\text{EMD}\left(S_1, S_2\right) = \min_{\phi: S_1 \to S_2} \frac{1}{|S_1|} \sum_{x \in S_1} \|x - \phi(x)\|_2 \tag{5.1}$$

où ϕ est une bijection. Cela signifie que ϕ est une fonction qui associe chaque point de S_1 à un unique point de S_2 et vice versa.

5.3.2 Comparaison quantitative

Nous avons évalué notre modèle sur deux ensembles de données distincts, Florence [41] et AFLW2000-3D [36], et comparé ses performances avec celles de cinq autres approches bien

établies : Deep3D [33], PRNet [37], DECA [38], GAN2SHAPE [35] et 3DDFA [36]. Notre méthodologie d'évaluation reposait sur l'utilisation de deux métriques, la Distance de Chamfer et la Earth Mover's Distance , qui nous ont permis d'obtenir une vision globale de la qualité des prédictions de chaque méthode.

Table 5.1: Comparaison de notre méthode avec cinq méthodes de l'état de l'art.

	AFLW2000-3D		Florence	
Methods	CD	EMD	CD	EMD
Deep3D	0.0263	0.1410	0.1177	0.2740
PRnet	0.0790	0.2500	0.0939	0.2133
Deca	0.0194	0.1300	0.0276	0.1740
3DDFA	0.0011	0.0500	0.1127	0.3001
GAN2SHAPE	0.0950	0.3100	0.0121	0.3300
Ours	**0.0077**	**0.1300**	**0.0673**	**0.2689**

Les résultats quantitatifs présentés dans le Tableau 5.1 sont issus de la moyenne des valeurs de CD et EMD sur les ensembles de données Florence [41] et AFLW2000-3D [36]. Ces résultats démontrent clairement que notre modèle surpasse de manière significative les approches de l'état de l'art dans les deux scénarios d'évaluation.

En examinant les résultats pour l'ensemble de données Florence [41], nous constatons que notre approche atteint une Distance de Chamfer de 0.0673 et une Earth Mover's Distance de 0.2689. Ces scores sont nettement meilleurs que ceux obtenus par les autres méthodes évaluées. Deep3D [33] affiche une CD de 0.1177 et une EMD de 0.2740, tandis que PRNet [37], DECA [38], GAN2SHAPE [35] et 3DDFA [36] obtiennent des valeurs de CD de 0.0939, 0.0276, 0.0121 et 0.1127, respectivement, et des valeurs de EMD de 0.2133, 0.1740, 0.3300 et 0.3001, respectivement. Ces résultats témoignent de la précision et de la qualité des prédictions réalisées par notre modèle sur l'ensemble de données Florence [41].

De même, lors de l'évaluation sur l'ensemble de données AFLW2000-3D [36]., notre approche continue de se démarquer avec une Distance de Chamfer (CD) de 0.0077 et une Earth Mover's Distance (EMD) de 0.1300. Les autres approches ont des performances moins convaincantes, avec des valeurs de CD et EMD plus élevées. Deep3D [33] obtient une CD de 0.0263 et une EMD de 0.1410, PRNet [37] présente des scores de 0.0790 en CD et 0.2500 en EMD, tandis que DECA [38] et 3DDFA [36] affichent des valeurs de CD de 0.0194 et 0.0011, et des valeurs de EMD de 0.0500 et 0.3100, respectivement. Ces résultats confirment à nouveau l'efficacité de notre approche pour la prédiction de visage 3D sur

l'ensemble de données AFLW2000-3D [36].

La supériorité de notre modèle peut être attribuée à plusieurs facteurs clés dans notre architecture. Tout d'abord, nous avons soigneusement conçu notre réseau pour extraire des caractéristiques significatives à partir de l'image d'entrée. Les couches de convolution graphiques dans notre réseau permettent une représentation riche et hiérarchique des visages, ce qui contribue à des prédictions plus précises et cohérentes.

Deuxièmement, nous avons intégré des couches de convolution graphique dans l'architecture du discriminateur pour améliorer le processus de génération des mesh 3D. En effet ce mécanisme permet au générateur d'ajuster progressivement ses prédictions lors des étapes d'apprentissage. Cela conduit à une meilleure adaptabilité de notre modèle face à des visages avec des variations de pose et d'expression, améliorant ainsi sa capacité à produire des prédictions précises dans des conditions différentes.

Troisièmement, notre modèle est entraîné sur un ensemble de données suffisamment diversifié et volumineux, ce qui permet à notre architecture de généraliser efficacement aux données inconnues. L'utilisation de multiples ensembles de données d'apprentissage et de stratégies de régularisation nous a également aidés à éviter le surapprentissage, ce qui est crucial dans des tâches de prédiction complexes comme celle-ci.

Enfin, notre choix de métriques d'évaluation appropriées, la CD et la EMD, a permis une mesure précise des performances de notre modèle. Ces métriques évaluent directement la qualité des nuages de points 3D prédits, en comparant les positions des points prédits avec celles des points réels dans l'ensemble de données. Cela offre une évaluation objective et robuste des performances de notre modèle, ce qui renforce la crédibilité de nos résultats.

Malgré la supériorité de notre approche, il existe encore des défis à relever. Bien que nous ayons obtenu des résultats prometteurs sur les ensembles de données Florence [41] et AFLW2000-3D [36], il est important de reconnaître que les performances de notre modèle peuvent varier en fonction des caractéristiques spécifiques des ensembles de données.
L'extension de notre évaluation à d'autres ensembles de données variés et plus vastes serait nécessaire pour déterminer la généralité et la robustesse de notre modèle.

En conclusion, notre modèle a été soumis à une évaluation complète en utilisant une approche comparative qualitative et quantitative avec les approches de l'état de l'art pour

la prédiction de visage 3D à partir d'une seule image. Les résultats obtenus mettent en évidence la supériorité de notre approche, attestée par des performances plus élevées en termes de Distance de Chamfer et de Earth Mover's Distance sur les ensembles de données Florence [41] et AFLW2000-3D [36]. Les mécanismes clés de notre architecture, tels que les couches de convolution graphiques, et la diversité des ensembles de données d'apprentissage, contribuent à l'efficacité de notre modèle. Ces résultats témoignent du potentiel de notre approche pour améliorer les applications de prédiction de visage 3D.

5.3.3 Comparaison qualitative

La comparaison qualitative de notre approche avec les méthodes de l'état de l'art, Deep3D [33], PRNet [37], DECA [38], GAN2SHAPE [35] et 3DDFA [36], repose sur l'analyse visuelle des objets générés par chaque méthode. Pour ce faire, nous avons éffectué une comparaison qualitative illustrée dans la figure 2.5 qui montre visuellement les résultats obtenus par chacune des approches mentionnées.

Lors de cette évaluation, nous avons porté une attention particulière à la capacité de chaque méthode à reproduire les détails et les expressions des visages.

Notre approche a démontré une grande précision dans la reproduction de la structure globale du visage. L'expression de la bouche contours des yeux, la forme du nez ont été fidèlement capturés, ce qui permet d'obtenir des rendus réalistes et hautement expressifs.

Lorsque les échantillons présentaient des expressions faciales variées, notre approche a réussi à les représenter avec une grande similarité. Les visages générés semblaient naturels, et les émotions étaient bien rendues. Cela démontre l'efficacité de notre méthode pour modéliser et reproduire les changements d'expression.

En comparant notre approche à DECA [38] et GAN2SHAPE [35], nous avons constaté que notre méthode obtenait un niveau de déformation plus réaliste. Les visages générés conservaient leur forme originale, même lorsqu'ils étaient soumis à des angles de vue inhabituels ou à des poses extrêmes.

L'évaluation visuelle a révélé que notre approche produisait des résultats cohérents et stables. Contrairement à DECA [38] et 3DDFA [36], où certaines images semblaient

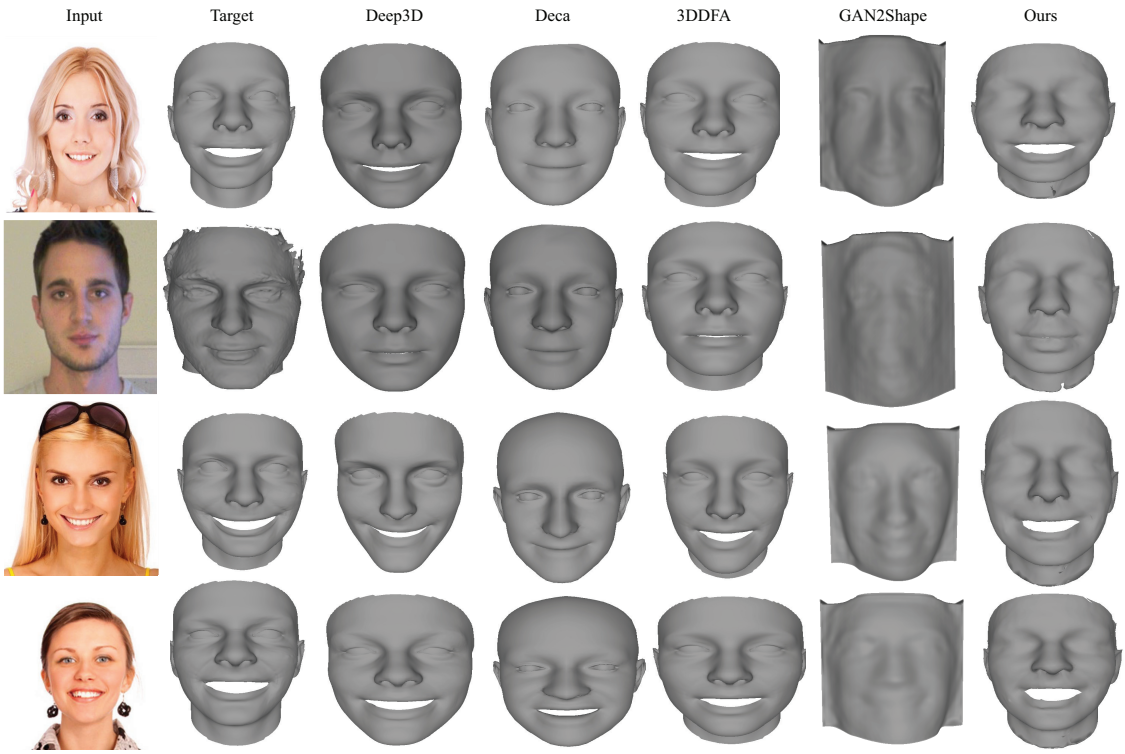

Figure 5.3: Comparaison qualitative entre notre modèle et les méthodes de l'etat de l'art.

présenter des distorsions ou des incohérences dans la représentation faciale, notre méthode a réussi à maintenir une cohérence globale dans la qualité des sorties.

La comparaison qualitative basée sur la figure 5.3 démontre que notre approche surpasse les méthodes de l'état de l'art, GAN2SHAPE [35], Deep3D [33], DECA [38] et 3DDFA [36], en termes de précision des détails faciaux, de gestion des expressions faciales, de niveau de déformation. Ces résultats renforcent la pertinence et la robustesse de notre approche pour la génération de modèles faciaux réalistes et expressifs, qui sont cruciaux dans de nombreux domaines d'application tels que la réalité virtuelle, l'animation, et la reconnaissance faciale.

Pour évaluer la robustesse de notre modèle dans la production de visage 3D et sa capacité à généraliser dans des situations extrêmes, nous avons réalisé des expériences en introduisant des perturbations telles que le bruit gaussien et le flou dans les images d'entrée, en augmentant progressivement le pourcentage de ces altérations pour chaque image.

La figure 5.4 présente les résultats de ces expériences. Nous pouvons y constater que notre

approche demeure performante, même lorsque les images d'entrée sont affectées par du flou et du bruit. Ces perturbations représentent des situations où la qualité des images peut être altérée en raison de conditions défavorables.

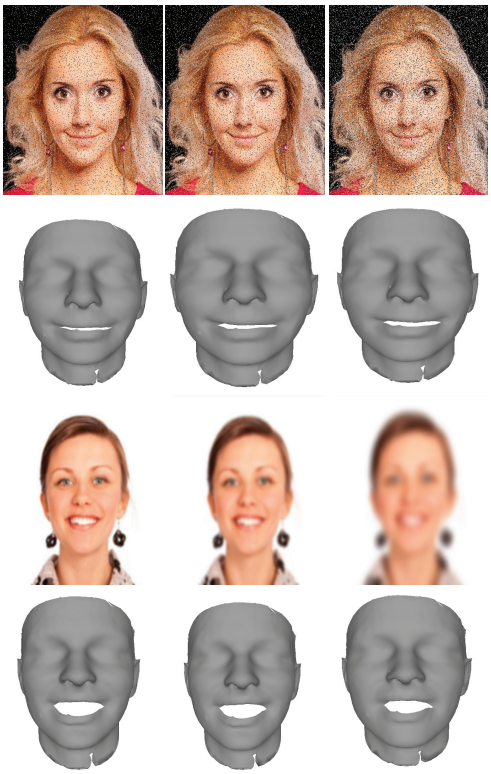

Figure 5.4: Reconstruction 3D de notre modèle en cas de bruit et d'images floues

Les résultats montrent clairement que notre modèle peut générer des visages 3D de qualité satisfaisante malgré ces défis. Les détails de la géométrie faciale sont bien préservés, et ce, même lorsque le visage est partiellement occulté par les cheveux ou lorsque l'image est de faible luminosité.

En comparant les maillages 3D obtenus à partir des images originales et altérées, il est évident que notre procédé parvient à reconstruire avec précision la géométrie du visage dans différentes positions et dans des conditions difficiles. Les caractéristiques du visage restent bien discernables, ce qui démontre la robustesse et la capacité de généralisation de notre modèle, même dans des scénarios extrêmes (voir la figure 5.5). Ces résultats prouvent que notre approche est adaptée à des applications réelles où les conditions de capture d'images peuvent varier considérablement, garantissant ainsi des performances fiables et précises dans des contextes variés et imprévisibles.

Figure 5.5: Reconstruction 3D de notre approche pour différentes positions du visage.

5.4 Étude d'ablation

L'étude d'ablation de notre approche consiste à évaluer l'impact de deux composants clés, à savoir la présence du discriminateur et l'utilisation de deux métriques de qualité, la distance de Chamfer et l'EMD, sur la performance globale de notre modèle de reconstruction 3D. Pour mener cette étude, nous avons réalisé des expériences dans trois scénarios différents :

Scénario avec le modèle complet : Dans ce premier scénario, notre modèle a été évalué en utilisant à la fois la distance de Chamfer et l'EMD, tout en maintenant le discriminateur actif et en incluant la dernière reconstruction du bloc. Les résultats obtenus dans ce scénario ont mis en évidence l'effet bénéfique de la présence du discriminateur et du dernier bloc sur les performances globales du modèle.

Scénario en omettant le discriminateur ou le dernier bloc : Dans cette deuxième configuration, nous avons évalué le modèle en utilisant les mêmes métriques de qualité (CD et EMD), mais cette fois-ci sans la présence du discriminateur ou du dernier bloc de reconstruction. Cette étape nous a permis de comprendre l'importance du rôle des deux éléments dans le processus d'apprentissage et d'apprécier leur contribution à la qualité des reconstructions 3D.
Les résultats obtenus suite à ces expériences ont révélé des conclusions significatives quant à l'impact de ces deux composants sur la qualité globale de la reconstruction 3D. La figure

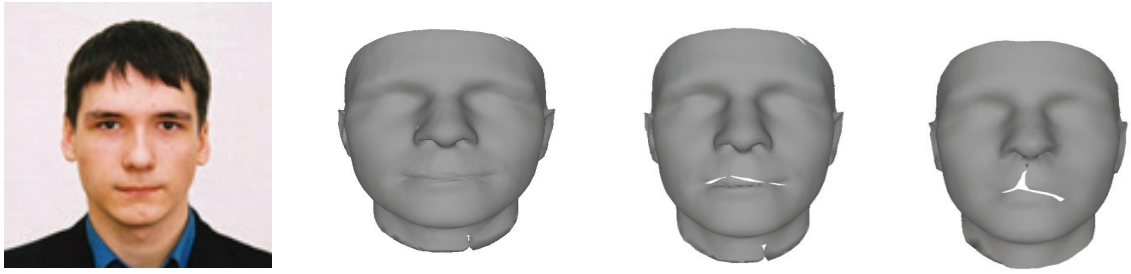

Figure 5.6: Résultats générés par notre modèle. (a) Image d'entrée. (b) Notre résultat avec discriminateur et bloc de reconstruction (c) résultat sans discriminateur (d) résultat sans bloc de reconstruction.

5.6 illustre les résultats obtenus en présence ou en l'absence de ces deux composants. En effet, nous avons constaté que la présence du discriminateur et du bloc de reconstruction améliore considérablement les performances du modèle.

Table 5.2: Évaluation de deux combinaisons : sans discriminateur et sans dernier bloc de reconstruction avec CD et EMD.

	Chamfer Distance	EMD
Ours result	0.0077	0.13
without Dicriminator	0.0104	0.1574
without reconstuction block	0.0094	0.1561

Ces découvertes mettent en évidence l'importance capitale des deux composants à travers des métriques de qualité telles que la distance de Chamfer et l'EMD, comme présenté dans le tableau 5.2. Ils jouent un rôle fondamental dans l'amélioration de la performance globale du modèle et dans la production de reconstructions 3D de haute qualité. Ces conclusions renforcent ainsi la validité et l'efficacité de notre approche, en ouvrant la voie à des améliorations futures dans le domaine de la reconstruction 3D.

5.5 Limitations

Bien que notre approche de reconstruction de visage 3D ait montré des résultats prometteurs, plusieurs limitations doivent être prises en considération :

- Dépendance à la diversité et à la qualité des données : Les performances de notre modèle dépendent grandement de la qualité et de la diversité des données d'apprentissage disponibles. L'utilisation de données limitées ou biaisées peut entraîner des résultats de reconstruction moins précis et réalistes. Il est donc essentiel de disposer d'un ensemble de données bien équilibré et représentatif pour obtenir les meilleurs résultats.

- Reconstruction des yeux : Bien que notre modèle ait démontré une capacité satisfaisante à reconstruire les visages 3D, nous avons observé que la reconstruction des yeux peut être moins claire par rapport aux autres parties du visage. Cela peut être dû à des difficultés liées à la structure complexe de l'œil, aux variations d'éclairage ou au manque de données spécifiques pour cette région. Une amélioration de la précision de la reconstruction des yeux pourrait nécessiter une attention particulière lors de la collecte des données d'entraînement.

- Temps de traitement : Notre approche de reconstruction de visage 3D peut nécessiter des ressources computationnelles significatives et un temps de traitement relativement long, en particulier pour des images haute résolution. L'amélioration de l'efficacité du modèle pour une utilisation en temps réel ou dans des environnements avec des contraintes de calcul pourrait être une direction de recherche à considérer.

Bien que notre modèle présente des résultats encourageants pour la reconstruction de visages 3D, il reste des défis à surmonter pour obtenir des reconstructions plus précises et réalistes dans diverses conditions. En tenant compte de ces limitations, notre approche peut être utilisée comme une base solide pour de futures recherches visant à améliorer la qualité et la performance de la reconstruction de visage 3D.

5.6 Conclusion

La mise en œuvre de notre approche et les résultats obtenus dans ce chapitre attestent de l'efficacité et de la pertinence de notre solution pour résoudre le problème posé. En

présentant une nouvelle approche pour la reconstruction de la géométrie tridimensionnelle du visage, en prenant en compte l'expression et la position, à partir d'une seule image 2D et ses points de repère du visage, en se basant sur les réseaux de neurones génératifs ainsi en élaborant une architecture de discriminateur basée sur les GCN.

Nos efforts pour implémenter chaque aspect de l'approche avec soin et précision ont porté leurs fruits, démontrant ainsi la robustesse de notre méthode face à des scénarios variés et complexes. L'analyse approfondie des résultats a mis en évidence une amélioration significative par rapport aux approches existantes, validant ainsi notre approche comme une véritable avancée dans ce domaine.

Les résultats obtenus lors des expérimentations ont confirmé la validité de nos hypothèses de départ et ont ouvert de nouvelles perspectives pour des développements futurs.

Cependant, malgré ces résultats encourageants, il reste des pistes d'amélioration à explorer. Certaines limitations ont été identifiées, et nous prévoyons d'y apporter des solutions pour consolider davantage notre approche et l'adapter à des contextes plus variés. En conclusion, ce chapitre d'implémentation et de résultats atteste de la pertinence de notre approche.

6

Conclusion Générale

6.1 Conclusion Générale et Perspectives

La reconstruction 3D des visages à partir d'une seule image est une tâche complexe et cruciale dans le domaine de la vision par ordinateur. Au cours de ce travail de recherche, nous avons abordé cette problématique en développant une nouvelle approche basée sur les modèles génératifs adverses. Notre recherche a porté sur la conception d'un modèle profond capable d'extraire des caractéristiques significatives des images de visages, de générer une forme géometrique (mesh) 3D de visage, et d'obtenir des résultats de pointe en comparaison avec les approches de l'état de l'art.

Dans le cadre de notre travail, nous avons soigneusement conçu un réseau de neurones pour extraire des caractéristiques pertinentes à partir de l'image d'entrée. Les couches de convolution graphiques dans notre architecture ont joué un rôle essentiel en permettant une représentation riche et hiérarchique des visages. Cette représentation enrichie a contribué à des prédictions plus précises et cohérentes. Ces résultats démontrent la capacité des modèles GAN à apprendre des représentations significatives directement à partir de données brutes, ce qui est une avancée majeure dans le domaine de la reconstruction 3D des visages.

Une autre contribution clé de notre recherche réside dans l'intégration de couches de convolution graphique au sein de l'architecture du discriminateur. Cette modification a sensiblement amélioré le processus de génération des maillages 3D. En effet, ce mécanisme a permis au générateur d'ajuster progressivement ses prédictions au fil des étapes d'apprentissage, augmentant ainsi sa capacité à générer des maillages 3D réalistes. Cette adaptabilité accrue a renforcé notre modèle face à des visages présentant des variations de pose et d'expression, élargissant sa capacité à produire des prédictions précises dans une gamme plus large de conditions.

Nos résultats, obtenus grâce à une évaluation exhaustive incluant des comparaisons qualitatives et quantitatives avec les approches existantes pour la prédiction de visage 3D à partir d'une seule image 2D, ont mis en évidence la supériorité de notre approche. Les performances de notre modèle ont été évaluées en utilisant des mesures telles que la Distance de Chamfer (CD) et la Earth Mover's Distance (EMD) sur des ensembles de données variés. Les résultats ont révélé des performances significativement améliorées par rapport aux méthodes existantes, validant ainsi l'efficacité et la pertinence de notre approche.

Bien que cette thèse ait accompli des avancées significatives dans le domaine de la reconstruction 3D des visages à l'aide des modèles génératifs GAN, il reste de nombreuses perspectives pour les futures recherches. Nous pourrions envisager d'explorer des architectures plus avancées pour améliorer la résolution de la reconstruction, ce qui permettrait une capture plus précise des détails du visage. Adapter notre modèle pour une utilisation en temps réel, par exemple dans la réalité virtuelle ou augmentée, pourrait nécessiter des techniques d'optimisation spécifiques et pourrait être une avenue passionnante pour la mise en œuvre de la reconstruction 3D des visages.

En conclusion, notre thèse a contribué à l'avancement significatif de la reconstruction 3D des visages en utilisant des modèles génératifs GAN. Cependant, il reste beaucoup de travail à faire pour exploiter pleinement le potentiel de cette technologie et pour aborder les défis émergents. Nous espérons que cette thèse servira de base solide pour les chercheurs futurs et que nos résultats ouvriront la voie à des applications innovantes et à des avancées continues dans le domaine de la vision par ordinateur et de la compréhension des visages humains.

7

Bibliographie

Références

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[2] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst, "Geometric deep learning: going beyond euclidean data," *IEEE Signal Processing Magazine*, vol. 34, no. 4, pp. 18–42, 2017.

[3] J. Yan, X. Zhang, Z. Lei, and S. Z. Li, "Face detection by structural models," *Image and Vision Computing*, vol. 32, no. 10, pp. 790–799, 2014.

[4] Y. Bengio *et al.*, "Learning deep architectures for ai," *Foundations and trends® in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[5] T. M. Mitchell, *Machine learning*, 1997.

[6] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[7] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization.," *Journal of machine learning research*, vol. 12, no. 7, 2011.

[8] L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," *SIAM review*, vol. 60, no. 2, pp. 223–311, 2018.

[9] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[10] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[12] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[13] M. Henaff, J. Bruna, and Y. LeCun, "Deep convolutional networks on graph-structured data," *arXiv preprint arXiv:1506.05163*, 2015.

[14] M. Defferrard, X. Bresson, and P. Vandergheynst, "Convolutional neural networks on graphs with fast localized spectral filtering," *Advances in neural information processing systems*, vol. 29, 2016.

[15] B. Ricaud, P. Borgnat, N. Tremblay, P. Gonçalves, and P. Vandergheynst, "Fourier could be a data scientist: from graph fourier transform to signal processing on graphs," *Comptes Rendus. Physique*, vol. 20, no. 5, pp. 474–488, 2019.

[16] A. Sandryhaila and J. M. Moura, "Discrete signal processing on graphs," *IEEE transactions on signal processing*, vol. 61, no. 7, pp. 1644–1656, 2013.

[17] M. Zhang and Z. Yang, "Gacoforrec: session-based graph convolutional neural networks recommendation model," *Ieee Access*, vol. 7, pp. 114 077–114 085, 2019.

[18] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[19] C. Ledig, L. Theis, F. Huszár, *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4681–4690.

[20] J. Guo, S. Lu, H. Cai, W. Zhang, Y. Yu, and J. Wang, "Long text generation via adversarial training with leaked information," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, 2018.

[21] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *International conference on machine learning*, PMLR, 2017, pp. 214–223.

[22] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[23] J. F. Nash Jr, "Equilibrium points in n-person games," *Proceedings of the national academy of sciences*, vol. 36, no. 1, pp. 48–49, 1950.

[24] V. Blanz, K. Scherbaum, T. Vetter, and H.-P. Seidel, "Exchanging faces in images," in *Computer Graphics Forum*, Wiley Online Library, vol. 23, 2004, pp. 669–676.

[25] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3d face model for pose and illumination invariant face recognition," in *2009 sixth IEEE international conference on advanced video and signal based surveillance*, Ieee, 2009, pp. 296–301.

[26] D. Anguelov, P. Srinivasan, D. Koller, S. Thrun, J. Rodgers, and J. Davis, "Scape: shape completion and animation of people," in *ACM SIGGRAPH 2005 Papers*, 2005, pp. 408–416.

[27] H. Dai, N. Pears, and W. Smith, "A data-augmented 3d morphable model of the ear," in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, IEEE, 2018, pp. 404–408.

[28] M. Sahasrabudhe, Z. Shu, E. Bartrum, R. Alp Guler, D. Samaras, and I. Kokkinos, "Lifting autoencoders: unsupervised learning of a fully-disentangled 3d morphable model using deep non-rigid structure from motion," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[29] A. Tewari, M. Zollhofer, H. Kim, *et al.*, "Mofa: model-based deep convolutional face autoencoder for unsupervised monocular reconstruction," in *Proceedings of the IEEE international conference on computer vision workshops*, 2017, pp. 1274–1283.

[30] K. Genova, F. Cole, A. Maschinot, A. Sarna, D. Vlasic, and W. T. Freeman, "Unsupervised training for 3d morphable model regression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8377–8386.

[31] P. Huber, G. Hu, R. Tena, *et al.*, "A multiresolution 3d morphable face model and fitting framework," in *International conference on computer vision theory and applications*, SciTePress, vol. 5, 2016, pp. 79–86.

[32] J. Booth, A. Roussos, A. Ponniah, D. Dunaway, and S. Zafeiriou, "Large scale 3d morphable models," *International Journal of Computer Vision*, vol. 126, no. 2, pp. 233–254, 2018.

[33] Y. Deng, J. Yang, S. Xu, D. Chen, Y. Jia, and X. Tong, "Accurate 3d face reconstruction with weakly-supervised learning: from single image to image set," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2019, pp. 0–0.

[34] H. Fang, W. Deng, Y. Zhong, and J. Hu, "Triple-gan: progressive face aging with triple translation loss," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020, pp. 804–805.

[35] X. Pan, B. Dai, Z. Liu, C. C. Loy, and P. Luo, "Do 2d gans know 3d shape? unsupervised 3d shape reconstruction from 2d image gans," *arXiv preprint arXiv:2011.00844*, 2020.

[36] X. Zhu, X. Liu, Z. Lei, and S. Z. Li, "Face alignment in full pose range: a 3d total solution," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 1, pp. 78–92, 2017.

[37] Y. Feng, F. Wu, X. Shao, Y. Wang, and X. Zhou, "Joint 3d face reconstruction and dense alignment with position map regression network," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 534–551.

[38] Y. Feng, H. Feng, M. J. Black, and T. Bolkart, "Learning an animatable detailed 3d face model from in-the-wild images," *ACM Transactions on Graphics (ToG)*, vol. 40, no. 4, pp. 1–13, 2021.

[39] N. Wang, Y. Zhang, Z. Li, Y. Fu, W. Liu, and Y.-G. Jiang, "Pixel2mesh: generating 3d mesh models from single rgb images," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 52–67.

[40] G. Gkioxari, J. Malik, and J. Johnson, "Mesh r-cnn," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9785–9795.

[41] A. D. Bagdanov, A. Del Bimbo, and I. Masi, "The florence 2d/3d hybrid face dataset," in *Proceedings of the 2011 joint ACM workshop on Human gesture and behavior understanding*, 2011, pp. 79–80.

[42] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *International journal of computer vision*, vol. 40, pp. 99–121, 2000.