Segmentation automatique d'images échocardiographiques à l'aide d'une architecture Shifted Windows Vision Transformer

par

Souha NEMRI

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

MONTRÉAL, LE "08/01/2025"

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

PRÉSENTATION DU JURY

CE MÉMOIRE A ÉTÉ ÉVALUÉ

PAR UN JURY COMPOSÉ DE:

M. Luc Duong, directeur de mémoire
Département de génie logiciel et des TI à l'École de technologie supérieure

M. Mohamed Cheriet, président du jury
Département de génie des systèmes à l'École de technologie supérieure

Mme. Sylvie Ratté, membre du jury
Département de génie logiciel et des TI à l'École de technologie supérieure

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

LE "04/12/2024"

À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

REMERCIEMENTS

Je tiens à exprimer mes sincères reconnaissances à mon professeur de recherche, Monsieur Luc Duong pour son encadrement tout au long de ces deux années. Je tiens à le remercier pour la confiance qu'il m'a accordée et pour l'opportunité qu'il m'a offerte afin d'entamer ma maîtrise à l'école de technologie supérieure. Ses vastes connaissances et son expertise m'ont guidé tout au long de la réalisation de mon mémoire. Sa patience et ses conseils m'ont permis de progresser avec confiance et de mener à bien mon projet avec rigueur et assurance. Monsieur Luc, était non seulement mon encadreur, mais encore une personne à mon écoute, toujours présente pour m'offrir son soutien et me faciliter le parcours. J'étais aussi chanceuse de l'avoir eu comme professeur.

Je témoigne aussi d'une sincère reconnaissance aux membres du jury, Monsieur Mohamed Cheriet et Madame Sylvie Ratté, d'avoir accepté de faire partie du jury et d'évaluer mon travail. Votre expertise et votre rigueur ont été d'une grande valeur dans la finalisation de ce mémoire. Je remercie également les cardiologues du CHU Sainte Justine, Dr Marie-Josée Raboisson et Dr Joaquim Miró pour leurs précieuses collaborations, ce qui a grandement contribué à enrichir mes connaissances des exigences cliniques. C'est ainsi que j'ai pu avoir une idée sur les besoins réels des patients, une étape essentielle à la pertinence de mon travail.

Comme étant membre de notre laboratoire LIVE, je voudrais remercier tous les étudiants que j'ai eu le plaisir de rencontrer au cours de mes études. Ce fut une expérience enrichissante de faire votre connaissance et de pouvoir partager des moments précieux ensemble, tant sur le plan académique que personnel.

Finalement, j'exprime ma gratitude à mes chers parents d'avoir cru en moi et de m'avoir apporté leur soutien moral, émotionnel et financier. Aucun hommage ne pourrait être digne de l'amour et de l'affection dont ils ne cessent de me combler. À mon frère et ma sœur, pour leur soutien et leurs encouragements constants. Même si la distance me sépare de ma famille, loin des yeux, près du cœur, vous êtes toujours présents dans mes pensées et dans mon cœur. À tous mes amis, je vous remercie également d'être là pour moi, d'avoir illuminé mes journées, même à l'étranger.

Je tiens à exprimer ma gratitude envers les Fonds de Conseil de recherches en sciences naturelles et en génie du Canada (CRSNG) pour leur soutien financier. Ce travail a également été soutenu par le programme Mitacs Globalink Graduate Fellowship. Leur appui a été essentiel à la réalisation de cette étude.

Segmentation automatique d'images échocardiographiques à l'aide d'une architecture Shifted Windows Vision Transformer

Souha NEMRI

RÉSUMÉ

L'échocardiographie est l'une des modalités d'imagerie les plus couramment utilisées pour le diagnostic des cardiopathies congénitales. L'analyse des images échocardiographiques est cruciale pour obtenir des informations précises sur l'anatomie cardiaque. Les modèles de segmentation sémantique peuvent être utilisés pour délimiter précisément les frontières du ventricule gauche et permettre une identification précise et automatique de la région d'intérêt, ce qui peut s'avérer extrêmement utile pour les cardiologues. Dans le domaine de la vision par ordinateur, les architectures de réseaux neuronaux convolutionnels (CNN) restent dominantes. Les approches CNN existantes se sont avérées très efficaces pour la segmentation de diverses images médicales au cours de la dernière décennie. Cependant, ces solutions ont généralement du mal à capturer les dépendances à longue portée, en particulier lorsqu'il s'agit d'images avec des objets de différentes échelles, de tailles variables et de structures complexes. Dans cette étude, nous présentons une méthode efficace de segmentation sémantique des images échocardiographiques qui surmonte ces défis en tirant parti du mécanisme d'auto-attention de l'architecture Transformer. Notre solution intègre un mécanisme d'attention, qui est une technique visant à se concentrer sur les éléments les plus pertinents d'une image. Cela contribue à une meilleure analyse des objets, de manière plus précise et plus efficace. Nous introduisons des modèles Shifted Windows Transformer models (Swin Transformers), qui encodent à la fois le contenu des structures anatomiques et les relations entre elles. Notre solution combine les architectures Swin Transformer et U-Net, tout en incorporant leurs avantages respectifs afin de renforcer les résultats. La validation de la méthode proposée est effectuée avec l'ensemble de données EchoNet-Dynamic utilisé pour entraîner notre modèle. Les résultats montrent une précision de 0,97, un coefficient de Dice de 0,87 et une Intersection over Union (IoU) de 0,78. C'est ainsi que les modèles de transformateur de Swin sont prometteurs pour la segmentation sémantique des images échocardiographiques en aidant les cardiologues à analyser et à évaluer automatiquement des images échocardiographiques complexes.

Mots-clés: Échocardiographie, segmentation sémantique, ventricule gauche, transformers, U-Net

Automatic segmentation of echocardiographic images using a Shifted Windows Vision Transformer architecture

Souha NEMRI

ABSTRACT

Echocardiography is one the most commonly used imaging modalities for the diagnosis of congenital heart disease. Echocardiographic image analysis is crucial to obtaining accurate cardiac anatomy information. Semantic segmentation models can be used to precisely delimit the borders of the left ventricle, and allow an accurate and automatic identification of the region of interest, which can be extremely useful for cardiologists. In the field of computer vision, convolutional neural network (CNN) architectures remain dominant. Existing CNN approaches have proved highly efficient for the segmentation of various medical images over the past decade. However, these solutions usually struggle to capture long-range dependencies, especially when it comes to images with objects of different scales and complex structures. In this study, we present an efficient method for semantic segmentation of echocardiographic images that overcomes these challenges by leveraging the self-attention mechanism of the Transformer architecture. The proposed solution extracts long-range dependencies and efficiently processes objects at different scales, improving performance in a variety of tasks. We introduce Shifted Windows Transformer models (Swin Transformers), which encode both the content of anatomical structures and the relationship between them. Our solution combines the Swin Transformer and U-Net architectures, producing a U-shaped variant. The validation of the proposed method is performed with the EchoNet-Dynamic dataset used to train our model. The results show an accuracy of 0.97, a Dice coefficient of 0.87, and an Intersection over Union (IoU) of 0.78. Swin Transformer models are promising for semantically segmenting echocardiographic images and may help assist cardiologists in automatically analyzing and measuring complex echocardiographic images.

Keywords: Echocardiography, Semantic segmentation, Left Ventricle, Transformers, U-Net

TABLE DES MATIÈRES

Page

LISTE DES FIGURES

Page

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

ETS	École de technologie supérieure
CNN	Convolutional neural network
SWA	Shifted Window Attention
ViT	Vision Transformer
ESV	Volume end-systolique
EDV	Volume end-diastolique

INTRODUCTION

L'échocardiographie est une technique d'imagerie la plus largement utilisée pour la visualisation des structures internes du cœur. Elle est largement utilisée par les cardiologues pour l'évaluation des maladies cardiaques grâce à sa capacité de fournir des données anatomiques en temps réel. L'une des applications de l'échocardiographie est l'analyse du ventricule gauche qui revêt une importance particulière. En effet, le ventricule gauche constitue la partie majeure du cœur et joue un rôle crucial dans la circulation sanguine en pompant le sang oxygéné pour fournir une alimentation dans le reste du cœur. C'est ainsi qu'une évaluation précise de la fonction et la structure du ventricule gauche est essentielle pour le diagnostic des pathologies telles que l'insuffisance cardiaque ou la cardiomyopathie.

Dans ce cadre, une segmentation précise des structures anatomiques est une tâche essentielle pour le traitement clinique de certaines maladies cardiaques. Certains paramètres cardiaques tels que les volumes de fin de systole et de fin de diastole, la fraction d'éjection sont de bons indicateurs de la santé cardiaque. En se basant sur ces paramètres, les cliniciens peuvent les utiliser pour la prise des décisions cliniques et le suivi des patients. Cependant, la segmentation automatique du ventricule gauche dans les images échocardiographiques pose de nombreux défis en raison de la variabilité des formes cardiaques et de la qualité des images obtenues. Jusqu'à présent, le cardiologue effectue manuellement les mesures directement à partir de l'image, un processus qui nécessite beaucoup de temps. Ainsi, le développement des méthodes automatiques pour la segmentation du ventricule gauche demeure un domaine de recherche actif et essentiel pour améliorer les soins cardiaques et la prise de décision.

La structure du ventricule gauche est complexe et varie considérablement d'un individu à un autre. En raison de la nature dynamique du cœur, les mouvements cardiaques et respiratoires d'un patient peuvent affecter significativement la qualité des images médicales . Cela nous confronte à un défi majeur qui pourrait entraîner une perte de la résolution spatiale et donc entraver la prise de décision clinique. En raison de la nature dynamique de notre organe,le cœur se déplace

d'une manière significative, ce qui rend l'acquisition des images échocardiographiques une problématique et engendre une mauvaise interprétation précise des structures anatomiques. La respiration naturelle entraîne aussi un déplacement du cœur de plus d'un centimètre à chaque cycle respiratoire, alors qu'un mouvement moyen de 8 mm se produit au cours des examens de tomographie par émission de positons (TEP) en raison de la respiration. L'acquisition d'images à haute résolution est rendue plus complexe par ces mouvements dynamiques, ce qui peut diminuer la fiabilité diagnostique des techniques d'imagerie telles que l'échocardiographie, l'IRM cardiaque ou la TEP (Rubeaux, Doris, Alessio & Slomka (2017)). C'est ainsi que le mouvement cardiaque engendre des défis importants pour maintenir la qualité des images, ce qui met en évidence l'importance cruciale de la segmentation dans l'analyse des données anatomiques, mettant ainsi en évidence la nécessité d'avoir recours à des techniques avancées pour garantir une interprétation précise. Auparavant, les recherches se sont orientées vers une grande variété de méthodes classiques de segmentation des images. Parmi ces techniques, on trouve les contours actifs (ou snakes), et les méthodes basées sur des régions. À titre d'exemple, les contours actifs ont une capacité de déterminer la région d'intérêt en fonction des discontinuités de luminosité dans l'image. On commence par initialiser une série de points mobiles dans la zone d'intérêt de l'image ou autour d'un objet. Et à chaque itération, on fait évoluer la disposition des points sous l'influence de forces internes et externes afin d'entourer les contours de la structure cible (Bi, Tan, Cheng, Chen & Wang (2022)).

Cependant, plusieurs modèles de contours actifs présentent des limitations lorsqu'il s'agit de segmenter avec précision des objets multiples qui se chevauchent. Dans ce cas, le modèle risque d'identifier plusieurs objets en un seul, ce qui entraîne une mauvaise séparation et une segmentation erronée. Leur incapacité à résoudre des objets multiples qui se chevauchent pourra influencer la précision. Ceci est particulièrement problématique dans la segmentation d'images médicales où la délimitation précise de plusieurs structures est essentielle (Fatakdawala *et al.* (2010)). De même, en présence de certaines conditions initiales, le processus de segmentation

peut échouer ou produire des résultats inexacts. Ces méthodes nécessitent un apprentissage manuel, où, à chaque passage ou itération, l'intervention humaine est nécessaire afin de réaliser certaines modifications pour ajuster les paramètres et corriger les erreurs. De même, l'une des limitations majeures des contours actifs est leur sensibilité aux structures concaves. Ces techniques s'appuient généralement sur l'identification des bords forts pour parvenir à une segmentation réussie. Cette dépendance ainsi que la formulation les rend moins efficaces dans les scénarios où les limites des objets ne sont pas bien définies ou sont obscurcies par le bruit. Cette contrainte de lissage a tendance à trop lisser les transitions nettes, ce qui entraîne une détection non précise des limites dans les régions où les contours devraient présenter des bords nets. La détection sera moins efficace dans les scénarios où les contours de notre région d'intérêt ne sont pas bien délimités. Dans ce cas, lors du traitement des images médicales, en particulier si on traite des structures concaves, les contours actifs peuvent donner une délimitation erronée (Hoang Ngan Le *et al.* (2020)).

Dans ce cadre, les approches basées sur l'apprentissage profond sont introduites afin de surmonter les limitations et les défis des techniques traditionnelles. Les réseaux de neurones convolutifs (CNN) ont révolutionné les méthodes traditionnelles de segmentation d'images médicales comme étant des solutions prometteuses. Dans le domaine de la vision par ordinateur, les CNN demeurent dominantes grâce à leur aptitude à apprendre les caractéristiques les plus complexes et pertinentes de l'image. Elles sont devenues l'élément central d'un grand nombre de tâches. Ces approches sont nettement plus puissantes que les anciennes théories en termes de précision et de robustesse. Ils se sont même révélés très efficaces pour la segmentation de diverses images médicales en s'affranchissant de la nécessité d'une intervention manuelle et en étant moins sensibles aux conditions initiales et au bruit. Néanmoins, les CNN présentent également plusieurs limitations en ce qui concerne l'extraction des dépendances globales à travers l'image. Les réseaux de neurones convolutifs (CNN) rencontrent des limites lorsqu'il s'agit de traiter des relations à longue portée, nécessitant ainsi de nombreuses couches pour obtenir une

compréhension contextuelle suffisante. Cette contrainte pourra engendrer une augmentation du coût de calcul et une complexité accrue du modèle.

Récemment, les Vision Transformer (ViT) ont connu un intérêt considérable et sont perçus comme une alternative prometteuse (Takahashi *et al.* (2024a)). Initialement, ils ont été développés pour des tâches de traitement du langage naturel tout en utilisant des mécanismes d'attention pour capturer des dépendances à longue portée dans les données. Le recours au CNN n'est même pas nécessaire et un ViT appliqué directement à des séquences d'images peut aboutir à de très bons résultats sur les images. Les ViT permettent de prendre en compte des relations globales dans l'image, offrant ainsi une meilleure compréhension contextuelle que les CNN traditionnels. Étant donné que les solutions basées sur le CNN ont généralement du mal à capturer les dépendances entre des éléments distants dans les images, en particulier lorsqu'il s'agit d'images avec des objets de différentes échelles et des structures complexes, notre choix s'est alors orienté vers le modèle Swin-Unet.

L'objectif principal de cette étude est de concevoir une méthode pour l'évaluation automatique de la fraction d'éjection. Une méthode basée sur Vision Transformer (ViT) est évaluée pour la segmentation du ventricule gauche à partir d'images échocardiographiques. Les ViT sont reconnus pour leur capacité à capturer les relations globales dans les images.

CHAPITRE 1

REVUE DE LITTÉRATURE

1.1 Introduction

Dans ce chapitre, nous examinerons les avancées récentes dans la segmentation du ventricule gauche à partir des images échocardiographiques. Dans un premier temps, nous exposerons l'anatomie du cœur tout en faisant une brève description de sa structure, de sa composition et de son principe de fonctionnement. Par la suite, nous examinerons les évolutions récentes dans les architectures de réseaux de neurones pour la vision par ordinateur, en mettant particulièrement l'accent sur l'émergence des ViT. Nous explorerons les avantages, les défis, et les applications potentielles de ces architectures.

1.2 Anatomie du cœur

Le cœur est un organe musculaire, structuré comme une pompe, dont la fonction principale est de propulser le sang à travers le système circulatoire (Magnin (2023)). Il est localisé au milieu de la poitrine d'un être humain, mais légèrement à gauche pour la plupart des gens. On peut le trouver à l'intérieur du médiastin, la cavité centrale du thorax. Le cœur permet d'assurer le déplacement du sang à travers les vaisseaux sanguins en le propulsant. Par des contractions régulièrement répétées, le cœur propulse chaque jour près de 8 000 litres de sang dans tous les organes du corps (Lavigne (2016)). Cette propagation du sang fait en sorte que le corps entier soit fourni en oxygène quand le cœur bat.

Par cette circulation sanguine, l'oxygène et les éléments nutritifs sont acheminés vers les différentes régions du corps, et le dioxyde de carbone et les déchets indésirables sont expulsés. Le cœur est donc l'organe principal de notre circulation sanguine. Afin de mieux préciser la circulation sanguine dans notre corps, on pourrait la subdiviser en deux grandes parties.

1.2.1 La circulation sanguine

Commençons par **la circulation pulmonaire**, dans ce cas, le cœur va recevoir le sang privé d'oxygène et riche en gaz carbonique. Il assure son transport vers les poumons, où il sera ré-oxygéné tout en éliminant l'excès de dioxyde de carbone. Cette circulation a lieu dans le côté droit du cœur (Beurnier (2021)).
Dans un deuxième temps, le côté gauche du cœur va recevoir du sang fraîchement oxygéné et pauvre en gaz carbonique en provenance des poumons. Il sera alors propulsé afin de jouer le rôle de transporteur d'oxygène, de nutriments dans tout le réseau de vaisseaux sanguins. Ce transport constitue **la circulation systémique** (Jessica I. Gupta (2022)).
Pour terminer, le sang regagne le cœur en parcourant les veinules, les veines et les grosses veines.

1.2.2 Les cavités du cœur

Le cœur se trouve protégé par une enveloppe à deux couches appelée péricarde. Une épaisse paroi divise le cœur en deux grandes parties. Sur le plan anatomique et fonctionnel, le cœur se compose de deux sections majeures, appelées également cœur droit et cœur gauche, qui renferment respectivement deux cavités distinctes. Dans la partie supérieure, on retrouve deux oreillettes, quant à la partie inférieure, elle renferme les ventricules. Ces chambres sont séparées par une paroi épaisse de tissu nommée le septum. Il est à noter que les ventricules constituent la partie majeure du cœur (Bittar (2022)).
Dans la partie supérieure, on retrouve **les oreillettes**, droite et gauche, aussi appelées atrium, lesquelles sont séparées par le septum inter-atrial. Cette cavité supérieure du cœur est destinée à réceptionner le sang de la circulation. Elles ont pour fonction de transmettre le sang des veines jusqu'aux ventricules puisqu'elles ne contribuent pas beaucoup à l'action de pompage du cœur.
Dans la partie inférieure, on observe **les ventricules** qui représentent la quasi-totalité de la masse du cœur. Cette zone constitue le point de départ du flux sanguin une fois que celui-ci a reçu des oreillettes. Ces quatre chambres communiquent entre elles pour garantir une circulation sanguine unidirectionnelle au sein du corps humain.

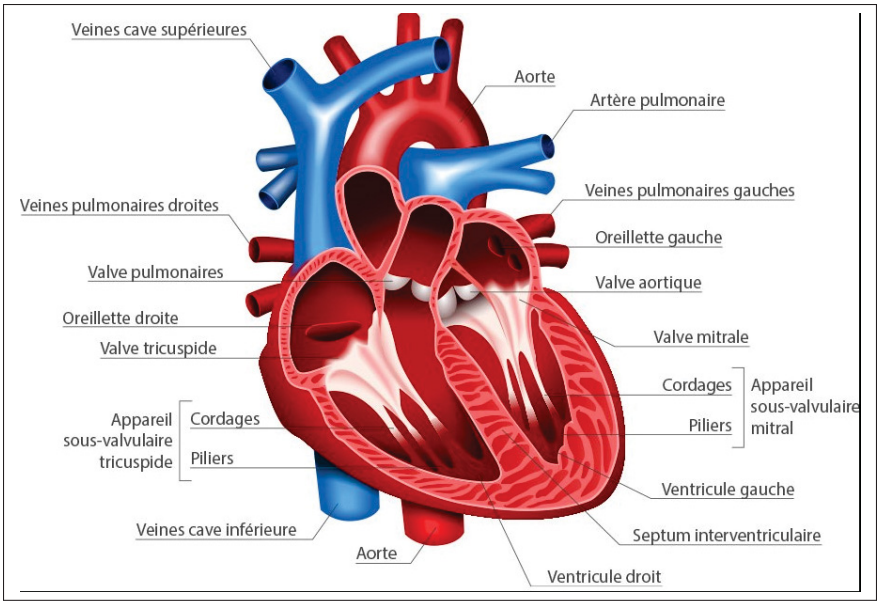

Figure 1.1 Structure anatomique du cœur
Tirée du site Santé sur le Net

Tout d'abord, l'oreillette droite accueille le sang appauvri en oxygène venant des différents organes. Une fois que l'oreillette droite est remplie, la valve tricuspide, en se contractant, achemine le sang au ventricule droit. Cette dernière, de son côté, va se contracter pour assurer la circulation du sang jusqu'aux poumons à travers la valve pulmonaire qui les relie. Dans ce cas, le rôle des poumons est de favoriser l'élimination du gaz carbonique pour avoir du sang oxygéné et l'envoyer vers l'oreillette gauche. Après contraction de cette dernière, la valve mitrale va à son tour s'ouvrir pour laisser couler le sang vers le ventricule gauche. À ce stade, le sang oxygéné est expulsé par la valve aortique vers l'aorte, qui va fournir une alimentation au reste du corps. Une fois que le sang riche en oxygène a parcouru tout le corps. Les veines se chargeront de reconduire le sang peu oxygéné vers l'oreillette droite, qui s'en remplira, et le cycle recommencera (Johanne Marcotte (2004)).

1.3 Échocardiographie

L'échocardiographie est une technique médicale utilisée pour obtenir des informations sur le cœur. Elle repose sur des ondes sonores pour générer des images permettant de visualiser notre organe.

Avec ces ultrasons, nous pouvons obtenir une représentation graphique des mouvements du cœur. Au cours de cet examen diagnostique, le médecin se servira de la sonde en la positionnant sur la poitrine de la personne pour produire des images des valves et des différentes chambres du cœur. De cette manière, le médecin est en mesure de visualiser le cœur, d'évaluer son action de pompage et de vérifier le fonctionnement des valves. En fonction des informations souhaitées par le médecin, le choix du type d'échocardiographie auquel le patient sera soumis peut varier (Michael J. Shea (2023)). Pour notre étude, on va se focaliser sur l'échocardiographie transthoracique qui se présente comme étant le type d'examen le plus couramment effectué dans le secteur médical. Dans ce cas, le cardiologue fait couvrir sa sonde avec un gel à ultrasons. En exerçant une légère poussée, des images de bonne qualité pourront être visualisées. Ainsi, les membres du corps médical sont en mesure de visualiser le cœur du patient, la dimension et la structure de ses quatre cavités, la pression sanguine, la fréquence de la circulation sanguine, les quatre valves cardiaques et les vaisseaux sanguins à proximité.

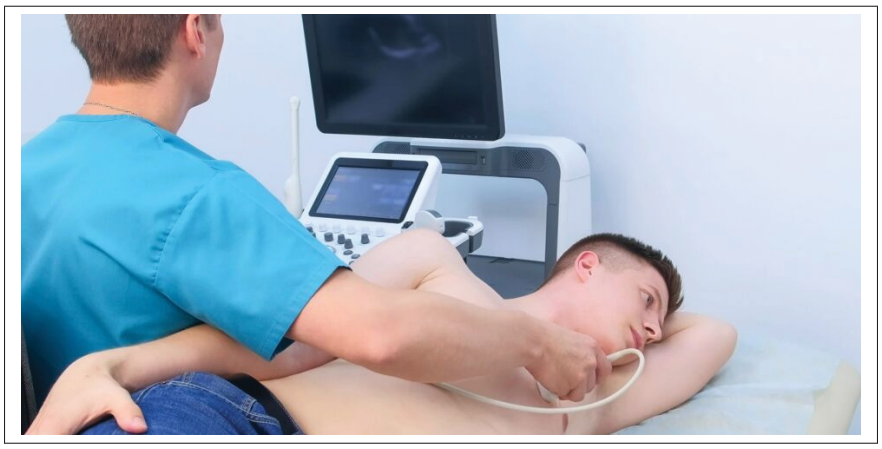

Figure 1.2 Échographie cardiaque
Tirée du site Femme Actuelle

1.4 Indice Clinique

1.4.1 Volume end-systolique (ESV)

Le volume end-systolique (ESV) est un indice clinique couramment utilisé en cardiologie. Il s'agit du volume du sang qui reste dans le ventricule à la fin de la phase systolique du cycle cardiaque. En d'autres termes, le cœur n'est pas en mesure de pomper tout le sang qui existe dans le ventricule durant la phase systolique. ESV représente alors le volume de sang le plus bas dans le ventricule à n'importe quel moment du cycle cardiaque.

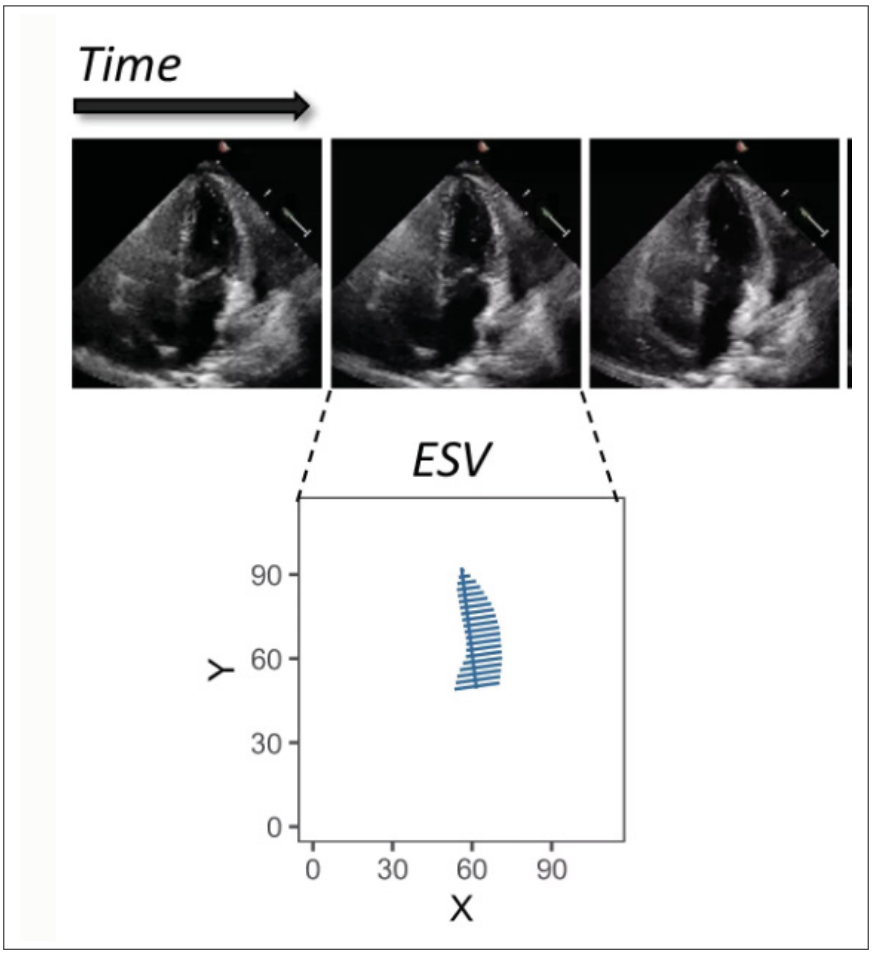

Figure 1.3 Volume end-systolique (ESV)
Tirée de l'article EchoNet-Dynamic

Il s'agit bien d'un indicateur important de la fonction cardiaque et de l'efficacité de la contraction

ventriculaire. Si on remarque que le volume ESV est plus élevé, il est possible que la contractilité cardiaque soit diminuée, ce qui est généralement le cas chez les patients qui présentent une insuffisance cardiaque ou une myocardiopathie dilatée. Cela signifie que le cœur n'expulse pas efficacement le sang pendant la contraction. Les cardiologues emploient généralement cet indicateur pour mesurer la sévérité du dysfonctionnement systolique, car l'ESV traduit clairement la capacité de contraction du ventricule. La mise en place d'un suivi de l'ESV permet de mieux orienter les décisions thérapeutiques et de modifier le traitement en tenant compte de l'évolution de la fonction cardiaque (Taboulet (2024)).
La figure 1.3 illustre le volume de fin de systole (ESV), qui est le volume minimal de sang présent dans le ventricule gauche à la fin de la contraction cardiaque. La partie supérieure de la figure montre une série d'images échocardiographiques représentant le mouvement du cœur pendant le cycle cardiaque, capturant le moment où le ventricule est le plus contracté. En bas, on trouve une visualisation graphique du volume de fin de systole, avec une représentation spatiale des contours du ventricule gauche. Ce type d'analyse est essentiel pour évaluer la fonction cardiaque, notamment en calculant des paramètres tels que la fraction d'éjection, un indice clinique de la performance du ventricule gauche.

1.4.2 Volume end-diastolique (EDV)

Le volume de fin de diastole (EDV) est le niveau de sang présent dans le ventricule gauche à la fin de la phase de remplissage (diastolique), juste avant la contraction du cœur. Les cardiologues considèrent généralement le EDV en tant qu'indicateur essentiel du fonctionnement cardiaque, puisque ce dernier quantifie la capacité du cœur à se remplir avant chaque contraction. Un EDV particulièrement élevé est souvent observé dans des conditions telles que l'insuffisance cardiaque congestive ou une surcharge volumique. Cependant, un faible EDV témoigne généralement du manque de capacité de remplissage du cœur, observé par exemple dans les cas de tamponnade cardiaque ou de cardiopathie restrictive. Généralement, les cardiologues ont recours à l'échocardiographie pour mesurer l'EDV en raison de son accessibilité et de son caractère non invasif. Ils pourront ainsi estimer avec précision les volumes ventriculaires et évaluer la fonction diastolique.

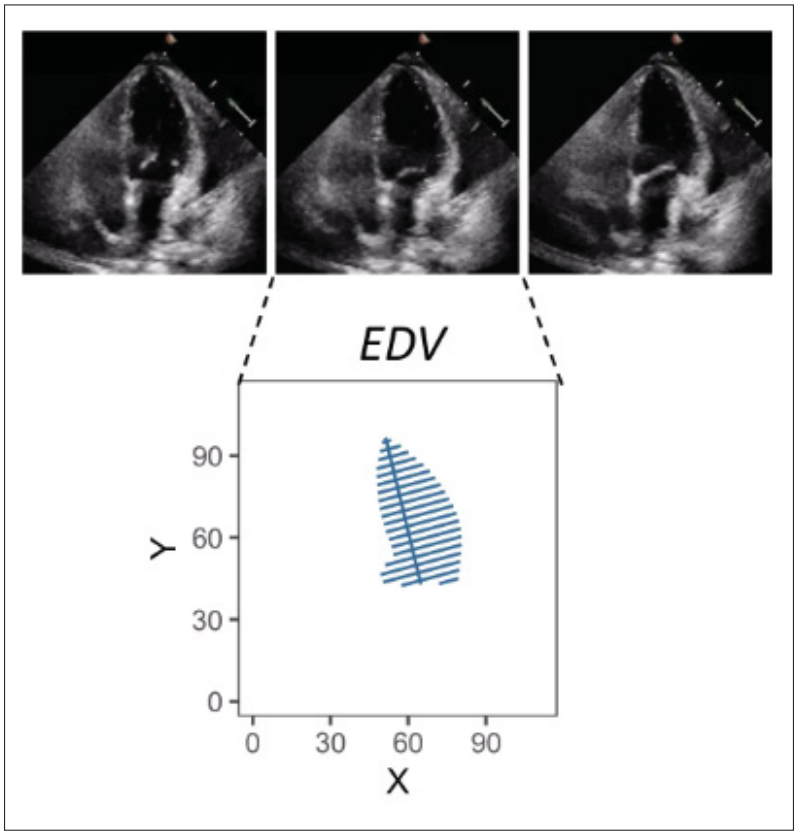

Figure 1.4 Volume end-diastolique (EDV)
Tirée de l'article EchoNet-Dynamic

Suite à ces mesures sur le patient, des interventions thérapeutiques sont prévues, notamment pour optimiser le traitement médical et la prise de décision concernant la pose de dispositifs. En incorporant la DV au processus global de diagnostic de la fonction cardiaque, les cardiologues pourront non seulement mesurer la gravité du dysfonctionnement ventriculaire, mais aussi contrôler de manière précise les effets des interventions thérapeutiques, ce qui permettra d'améliorer le pronostic et la qualité de vie des patients (Eske (2019)).

La figure 1.4 montre le volume end-diastolique (EDV), un indicateur essentiel de l'efficacité du cœur. Ci-dessus, une série d'images échographiques montrant le ventricule gauche lorsqu'il a atteint son volume sanguin maximal, juste avant que le cœur ne se contracte. Le graphique

montre les contours du ventricule dans un système de coordonnées cartésiennes, ce qui permet d'observer la géométrie et le volume précis du ventricule gauche à ce moment crucial.

1.4.3 Fraction d'éjection

Après avoir procédé à la réalisation du type d'échocardiographie le plus approprié, le médecin devra effectuer une lecture des valeurs permettant de mesurer la capacité du cœur du patient à pomper du sang enrichi en oxygène dans l'organisme. À cet égard, on pourra faire appel à **la fraction d'éjection** comme étant la plus importante variable. Elle se réfère à la capacité du cœur à expulser le sang des chambres inférieures du cœur (ventricules) à chaque fois qu'il se contracte. En général, pour un individu en bonne santé, cette fraction est un nombre élevé dont la valeur se trouve entre 50% et 70%. Une valeur basse est le signe que le cœur a des difficultés à satisfaire les nécessités de l'organisme (Eske (2024)).

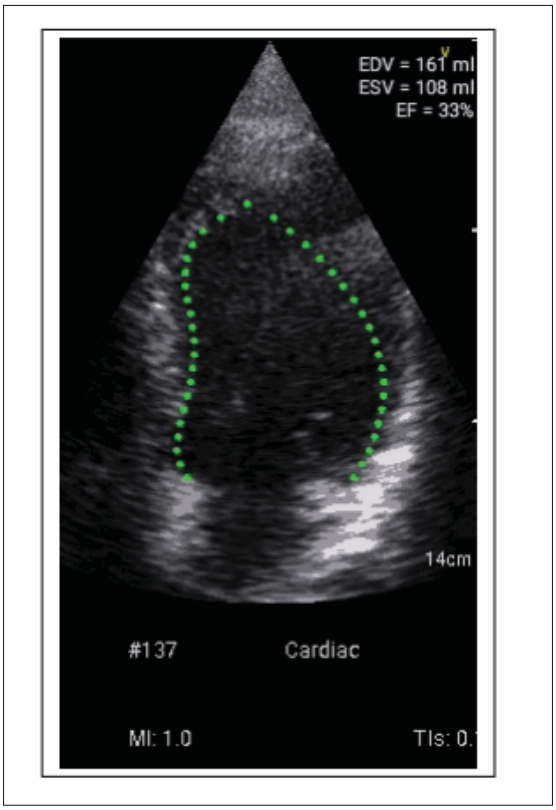

Figure 1.5 Calcul de la fraction d'éjection
Tirée du site Echocardiography

En examinant la relation entre le VDE, le VES et la FE, les cardiologues sont en mesure de déterminer la sévérité du dysfonctionnement cardiaque tout en adaptant les procédures thérapeutiques nécessaires. On peut procéder à un calcul simple pour déterminer la valeur de la fraction d'éjection. La formule est donnée par :

$$EF = \frac{EDV - ESV}{EDV} \tag{1.1}$$

En regroupant ces volumes pour calculer la FE, les cardiologues se dotent d'un outil performant pour mesurer la fonction cardiaque globale, améliorer les méthodes de soins et prédire le diagnostic des patients atteints d'une maladie cardiaque. Cette corrélation entre le EDV, le ESV et la FE est donc indispensable dans la prise de décision.

1.5 Analyse des séquences d'images échocardiographiques

La méthode de speckle tracking était très fréquemment employée dans le domaine de la cardiologie pour le suivi des images échocardiographiques. Cette méthode consiste à analyser le mouvement des tissus dans le cœur en utilisant les motifs granulaires (connus sous le nom de "speckles") présents dans les images échographiques. Dans ce cas, à chaque mouvement des tissus dans le cœur, on aura une représentation du trajet entre certains points qui permet de mesurer des mouvements de structures comme le muscle cardiaque.

Cela est aussi particulièrement utilisé pour évaluer certains paramètres cliniques et pour diagnostiquer tout dysfonctionnement cardiaque subtil. C'est ainsi que certaines études ont voulu étendre la fonctionnalité du speckle tracking pour l'adapter aux données 3D (Meunier (1998)). Alors cette méthode permet de suivre précisément les mouvements des tissus. Le suivi des mouvements dans les images échocardiographiques est une tâche cruciale en raison de la nature du speckle et des propriétés physiques de l'acquisition des images (Touil, Basarab, Delachartre, Bernard & Friboulet (2010)). Dans le cadre de la recherche sur les techniques de suivi du mouvement en imagerie échographique, une méthode se présente en exploitant l'influence de la décorrélation et de la géométrie sectorielle sur la précision du résultat à travers des simulations

réalistes. Concernant la géométrie sectorielle, elle fait référence à la configuration en forme de secteur utilisée dans les systèmes d'imagerie échographique, où les images sont obtenues sous forme de tranches ou d'angles définis autour d'un point central. La décorrélation désigne la réduction de la dépendance spatiale entre les pixels, ce qui permet d'optimiser l'extraction d'informations pertinentes. Cette méthode offre une nouvelle perspective pour le suivi des mouvements dans les images échocardiographiques. Cependant, il est nécessaire d'avoir recours à des méthodes plus avancées pour capter les relations complexes entre les pixels.

1.6 Vision Transformer (ViT)

Dans le domaine de la vision par ordinateur, les architectures de réseaux de neurones convolutifs (CNN) ont dominé la majorité des tâches telles que la segmentation sémantique et la détection d'objets. Leur aptitude à exploiter efficacement les cartes de caractéristiques à l'aide des opérations de convolution fait en sorte que ces modèles sont désormais parmi les architectures les plus performantes dans ce domaine. Cependant, malgré leur succès, les CNN présentent également des limitations, notamment leur faiblesse en capturant les relations globales entre les différentes parties d'une image, en particulier lorsqu'il s'agit des objets de différentes échelles et des structures complexes.

Dans ce contexte, les ViT ont récemment émergé comme étant une alternative prometteuse en raison de leur efficacité de calcul et de leur évolutivité. Initialement, ces architectures ont été employées pour des tâches de traitement du langage naturel tout en se basant sur des mécanismes d'attention. Cette approche a été appliquée aux images permettant de capturer les dépendances globales sans être restreinte par les fenêtres convolutives locales. C'est ainsi que ViT peut potentiellement surpasser les CNN dans certaines tâches tout en requérant moins de ressources informatiques et en aboutissant à des résultats performants.

1.6.1 L'évolution des modèles de vision : des CNN aux ViT

Différentes méthodes se présentent pour la segmentation du ventricule gauche et ont fourni des résultats satisfaisants. Les architectures de réseaux de neurones convolutifs ont été essentiellement

employées dans la majorité de ces études, offrant ainsi des solutions solides et performantes. Les architectures CNN, telles que les Fully Convolutional Networks (FCN) et les U-Net, sont particulièrement efficaces pour cette tâche. Certaines techniques ont largement été influencées par les méthodes manuelles en cherchant à reproduire la méthode employée par les spécialistes pour analyser des échantillons d'images médicales. La division de l'image en vue à axe court en plusieurs sections 2D est une technique couramment utilisée, ce qui permet de faciliter la segmentation de chaque zone individuellement. Pour effectuer cette tâche, certains travaux ont fait appel à FCN en employant des poids pré-entraînés de VGG16 entraînés sur la base de données ImageNet. Le modèle FCN-all-at-once-VGG16 fait intégrer des connexions résiduelles pour combiner des caractéristiques hiérarchiques issues de couches convolutionnelles à diverses échelles. Dans l'intégration de ces connexions, le modèle réussit à combiner des prédictions obtenues à des résolutions différentes (8, 16 et 32 pixels), ce qui améliore la précision spatiale des segments générés. Néanmoins, bien que cette méthode ait pu entraîner des résultats similaires aux données de segmentation manuelle, avec une sensibilité et des scores de similarité élevés, elle comporte des contraintes. La manipulation manuelle de l'utilisateur sur les images diminue la performance du modèle. Le fonctionnement de l'algorithme n'est donc pas entièrement automatisé, et sa précision dépend de la qualité de la première étape de traitement (Koo *et al.* (2020)). Bien que l'architecture traditionnelle d'un CNN ait pu avoir de bonnes performances pour la segmentation du ventricule gauche, elle montre des limites lors du traitement des zones pathologiques. La problématique repose également au niveau de la gestion des détails fins de l'image médicale, qui peuvent être difficiles à identifier et à analyser correctement. De ce fait, afin de faire face à ces variations significatives des organes, que ce soit en termes de taille ou de forme, entre les patients, il est possible d'apporter des améliorations au modèle Unet en intégrant un mécanisme attention gate. Dans cette situation, l'architecture finale Attention U-Net utilisant une pyramide d'images sert à identifier les informations spatiales pertinentes à partir de la carte des caractéristiques en utilisant des Attention Gates (AG). L'objectif est d'ajuster la contribution de chaque région dans la segmentation finale en tenant compte de sa fonction de pertinence. Ces AG produisent un coefficient d'attention alpha α compris entre 0 et 1 pour chaque pixel i. Une fois que les coefficients d'attention sont calculés, ils sont appliqués aux

cartes de caractéristiques afin de générer des caractéristiques ajustées qui incluent à la fois des informations contextuelles et des détails spatiaux correspondants. Ces caractéristiques ajustées sont ensuite propagées vers le décodeur pour aider à la reconstruction précise des segments. En résumé, l'objectif principal est de faciliter la gestion de la segmentation de petits objets complexes tout en préservant les informations spatiales requises (Cui, Yuwen, Jiang, Xia & Zhang (2021)). Bien que les différentes variantes de CNN ont pu démontrer une efficacité lors de la tâche de segmentation des images médicales tout aboutissant à des performances intéressantes, elles présentent également certaines limitations. La présence de fenêtres de réceptivité fixes dans les architectures CNN engendre une difficulté à capturer les relations à longue distance entre différentes parties de l'image, ce qui est essentiel pour une compréhension complète de sa structure globale (De Santi *et al.* (2023)). C'est ainsi que les ViT ont émergé comme une alternative novatrice aux CNN. Les Transformers ont été introduits dans le domaine du traitement du langage naturel (NLP). Inspirées par les performances de ViT, les études commencent à les intégrer directement aux images (Dosovitskiy *et al.* (2021)). Ces derniers ont démontré leur efficacité dans des tâches complexes, comme la segmentation ou la classification d'images médicales. À la comparaison des CNN qui se focalisent sur les relations locales à l'aide de filtres de convolution, les ViT ont la capacité de capturer des dépendances globales au sein d'une image. De même, la généralisation des réseaux de neurones profonds à une population plus large est un défi important. Notamment dans le domaine médical, si on se retrouve face à une grande base de données, ces modèles peuvent présenter des performances inégales. Bien que les CNN ont pu montrer de bons résultats plus précisément dans la détection des anomalies ECG, on se retrouve face à une inégalité de performance selon les ensembles de données utilisés pour l'entraînement. Cela limite leur application dans des situations réelles où on se retrouve face à des données variées. C'est ainsi que les ViT se présentent afin d'améliorer la généralisabilité en utilisant de grandes quantités de données. On pourra ainsi distinguer 3 modèles de tailles différentes : ViT-Base, ViT-Large et ViT-Huge. Grâce à l'utilisation d'apprentissage auto-supervisé, tels que les Masked Autoencoders (MAEs), les ViT pourront être utilisés pour des données limitées afin de reconstruire des informations manquantes. Cette méthode apprend à partir de données limitées, rendant le modèle plus adaptable et efficace face à des cas non vus. Cette étude est

généralement utilisée surtout dans le domaine médical où l'obtention de grandes quantités de données est souvent difficile en raison des coûts d'étiquetage et de l'approbation éthique (Takahashi *et al.* (2024b)).

1.6.2 Architecture de ViT

Pour une meilleure compréhension du fonctionnement de ViT, nous allons d'abord étudier son architecture globale afin de comprendre le rôle joué de chacun par ses composants.

- **Première étape :**

  ViT va diviser notre image d'entrée en plusieurs régions de tailles fixes. Le but de cette étape est de pouvoir structurer les données d'images d'entrées d'une façon qui ressemble à la structure des données dans le domaine de NLP.

$$g(x) = \frac{H * W}{P^2} \tag{1.2}$$

  L'équation ci-dessus illustre l'exemple que nous souhaitons définir. Prenons comme exemple une image de dimensions (H, W, N), où H représente la hauteur, W la largeur, N le nombre de canaux, et chaque segment possède une résolution de (P, P) pixels.

Figure 1.6 Résultat de la division de l'image
Tirée du site Implementing Vision Transformer (ViT) in PyTorch

Cette image est découpée en plus petites régions bidimensionnelles. Une fois les segments de l'image créés, ils seront convertis en une représentation unidimensionnelle en 1D. L'objectif de cette étape est de faire introduire ce résultat dans l'architecture Transformer pour un traitement ultérieur. L'aplatissement des régions constitue une phase cruciale de l'architecture

ViT, puisqu'elle permet au réseau de considérer l'image entière comme une séquence de jetons, et non comme une grille de valeurs de pixels. Ceci autorise l'architecture Transformer à récupérer efficacement les relations spatiales entre les différents segments, ce qui constitue un aspect majeur pour les tâches telles que la classification d'images et la détection d'objets.

- **Deuxième étape :**

  Par la suite, des encastrements positionnels vont être ajoutés afin d'intégrer des informations sur la position relative de chaque élément de notre image principale. Les embeddings positionnels sont des vecteurs supplémentaires indiquant la position de chaque patch dans notre séquence. Sans ces informations, le modèle aurait plus de mal à comprendre l'ordre et la relation spatiale entre les différentes parties de l'image, rendant la segmentation et la reconnaissance plus difficiles.

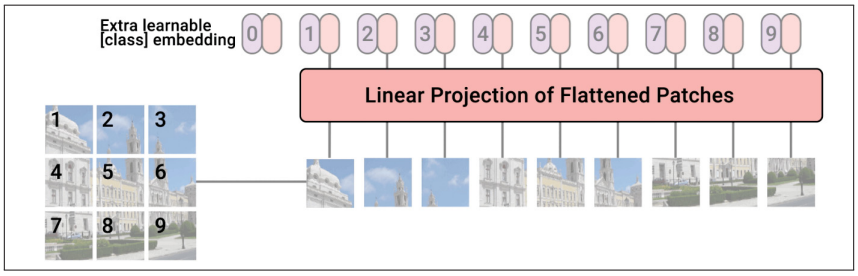

Figure 1.7 Résultat de la division de l'image
Tirée du site Implementing Vision Transformer (ViT) in PyTorch

La concaténation entre les incorporations positionnelles et linéaires produites à l'étape précédente est traitée par le mécanisme d'auto-attention dans l'architecture Transformer.

- **Troisième étape :**

  Dans cette étape, les encastrements positionnels et linéaires sont introduits dans un réseau de neurones profond nommé encodeur transformateur standard. Ce dernier emploie un mécanisme d'auto-attention pour le traitement de la séquence d'entrée et la détection des relations importantes qui existent entre les différents mots de la séquence.

Le mécanisme d'auto-attention permet au réseau de se focaliser sur les éléments les plus significatifs de la séquence et de donner un poids adéquat à leur importance.

Une fois la séquence d'entrée traitée, l'encodeur du modèle ViT génère une représentation de haut niveau de la séquence d'entrée, nommée séquence de sortie. Cette dernière est alors employée pour effectuer des prédictions sur l'image, comme la classification d'objets dans l'image, l'estimation de leur position et de leurs attributs, ou toute autre tâche souhaitée.

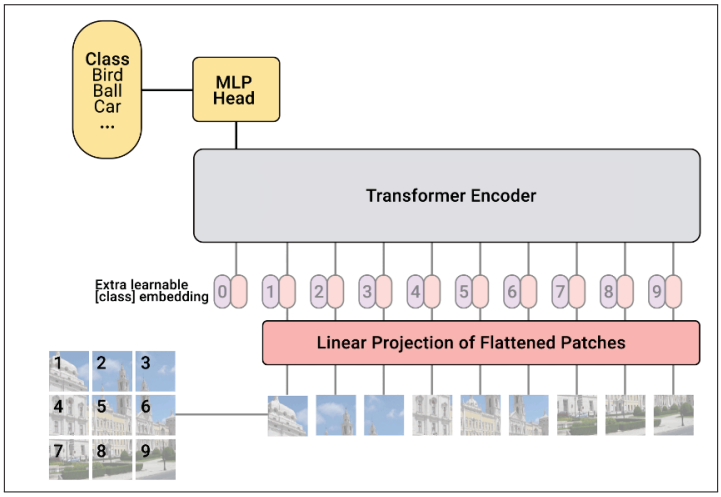

Figure 1.8 Fonctionnement du ViT
Tirée de l'article An Image is Worth 16x16 Words:
Transformers for Image Recognition at Scale

1.7 État de l'art des ViT

Dans le cadre du domaine de l'imagerie médicale, certains travaux se sont penchés sur la mise en œuvre de l'architecture ViT pour la segmentation du ventricule gauche. Ces études démontrent que la mise en commun de l'architecture d'un codeur-décodeur CNN et de la structure d'un ViT offre la capacité à exploiter à la fois les fonctionnalités d'apprentissage globales des transformers et les spécificités des réseaux Unet (Liu, He & Lu (2022)).

Au début, le CNN encodeur va extraire les caractéristiques de haut niveau des images échocardiographiques d'entrée. Son rôle est de capturer les différentes relations entre les motifs du sang,

le flux sanguin, afin de faciliter la tâche de segmentation du ventricule gauche.
Le résultat de cette étape sera introduit comme étant les composantes du réseau suivant, qui est le ViT. Son rôle est de détecter les caractéristiques sous-jacentes de l'image (comme les formes, les textures, les couleurs et les motifs, etc.) et de les convertir en une autre représentation en gardant seulement les attributs pertinents et de diminuer la dimensionnalité pour faire des prédictions sur la segmentation du ventricule gauche. À ce propos, le ViT utilise le mécanisme d'auto-attention pour pondérer l'importance de chaque pixel en fonction des relations avec les autres. Vers la fin, le CNN décodeur transforme les cartes caractéristiques à leur taille originale pour la classification finale par pixel. On pourra clairement voir que le modèle TransBridge proposé est désormais comme une nouvelle architecture qui a pu aboutir à une grande performance et une précision élevée dans la tâche de la segmentation, en comparant aux méthodes basées sur CNN seulement. Néanmoins, cette étude présente également certaines limitations. Le modèle utilisé repose principalement sur une architecture ViT, qui se caractérise par sa complexité. Le temps de calcul requis pour cette étude rend son adoption aux applications en temps réel et cliniques beaucoup plus difficile.

1.8 SegFormer

Dans cette section, nous allons aborder une autre architecture de réseau de neurones nommée SegFormer (Xie *et al.* (2021)). Il s'agit d'un nouveau concept dans le but de le concevoir pour les tâches de segmentation sémantique.
En effet, malgré le succès connu par les ViT ainsi que leur popularité dans plusieurs applications du domaine NLP ou vision par ordinateur comme la segmentation, la reconnaissance d'images ou la détection, on se retrouve face à plusieurs limitations. Dans le cas où nous abordons des images de large résolution, ViT nécessite une grande quantité de ressources informatiques en raison du grand nombre de paramètres du mécanisme d'auto-attention. En effet, les images d'entrée sont divisées en différentes régions et passent par plusieurs couches de l'architecture du réseau. Ce qui en résulte est un ensemble vaste de multiplications matricielles et donc une augmentation significative des coûts de calcul. Dans ce cas, on risque de dépasser les limites,

notamment sur les appareils disposant de ressources limitées. Afin de réduire ces contraintes, des techniques diverses ont été envisagées, notamment le recours à des mécanismes d'attention efficaces, la diminution du nombre de couches de ViT et l'utilisation d'incorporations de patchs hiérarchiques. Ces méthodes sont destinées à abaisser le coût de calcul de ViT tout en conservant son aptitude à traiter les images.

1.8.1 Architecture de SegFormer

L'architecture du modèle Segformer est composée de deux principaux modules : **A hierarchical Transformer encoder** qui est destiné à faire extraire les caractéristiques grossières et fines de notre image d'entrée ainsi qu'un **lightweight All-MLP decoder** pour fusionner directement ces caractéristiques multi-niveaux et prédire la segmentation sémantique.

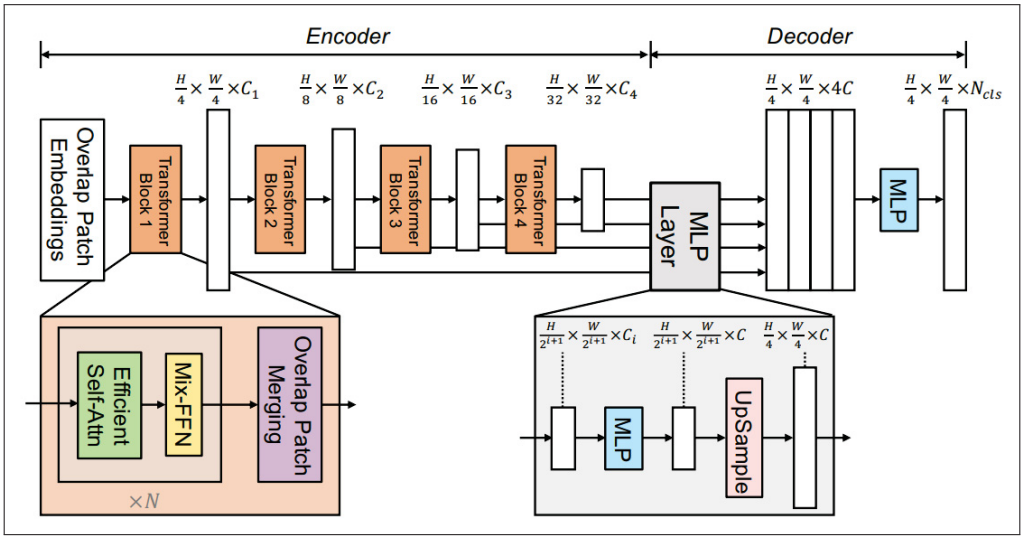

Figure 1.9 Architecture du modèle SegFormer
Tirée de l'article SegFormer: Simple and Efficient Design
for Semantic Segmentation with Transformers

1. **Hierarchical Transformer Encoder :**

   Il s'agit d'une version modifiée du Transformer encoder original utilisé dans les ViT. C'est un composant de réseau neuronal conçu spécialement pour traiter les caractéristiques de l'image multi-niveaux et multi-échelles. Ces caractéristiques fournissent à la fois des

informations grossières à haute résolution et fines à basse résolution, d'où l'amélioration de la performance de la segmentation sémantique contrairement au ViT.

Au début, le Transformer encoder hiérarchique va commencer par traiter les caractéristiques de l'image avec une série de couches convolutionnelles pour extraire les informations de bas niveau.

Ces derniers vont ensuite être introduits dans une succession de blocs de Transformer pour traiter les caractéristiques de haut niveau, puis ils vont passer à une couche de pooling pour réduire les dimensions spatiales. Et le cycle recommence, où les caractéristiques sont regroupées pour être traitées dans un autre ensemble de blocs Transformer encoder qui vont à leur part traiter les informations avec un niveau d'abstraction encore plus élevé.

C'est ainsi que l'architecture hiérarchique permet au modèle SegFormer de détecter les caractéristiques de l'image à plusieurs niveaux et de saisir à la fois les détails fins et les informations de haut niveau. Le traitement à plusieurs niveaux a permis au SegFormer d'extraire des caractéristiques plus significatives pour la segmentation et donc une meilleure performance et une grande précision.

2. **Lightweight All-MLP Decoder :**

Il s'agit d'un composant clé de l'architecture SegFormer dont le rôle est de générer la segmentation finale en se basant sur les caractéristiques extraites par le Transformer encoder et qui améliore l'efficacité et l'efficience de l'opération de décodage dans le modèle. Généralement, pour le processus de segmentation sémantique, le but d'utiliser un décodeur est d'augmenter la résolution spatiale des caractéristiques pour générer une sortie ayant les mêmes dimensions spatiales que l'image d'entrée. Ceci est effectué grâce à la présence de plusieurs couches d'upsampling.

Pour les architectures basées sur les CNN, la segmentation finale est générée par une série de couches convolutionnelles, qui impliquent généralement de nombreux paramètres apprenables et des coûts de calcul importants. Cela peut conduire à un surapprentissage, à une consommation de mémoire plus élevée et à des temps d'apprentissage plus longs. Mais dans cet article, les auteurs vont remplacer le décodeur du modèle CNN par All-MLP Decoder dont l'architecture est beaucoup plus efficace et plus performante dans les applications de

segmentation. ALL-MP remplace les couches convolutionnelles par des fully connected neural networks.

Le décodeur All-MLP léger est composé d'une série de modules de perceptron multicouche, qui sont de petits réseaux neuronaux entièrement connectés utilisés pour générer la carte de segmentation finale. Ces derniers sont beaucoup plus efficaces avec moins de paramètres et un coût de calcul plus faible par rapport aux couches convolutionnelles traditionnelles. C'est ainsi qu'on aura vers la fin une segmentation avec moins de ressources informatiques tout en réalisant de bonnes performances.

1.9 Swin Transformer

Pour les tâches de détection et de classification des images, les architectures de ViT connaissent un succès croissant. ViT a été conçu spécialement pour le traitement du langage naturel, mais on pourra clairement voir son implication dans plusieurs applications de vision par ordinateur et plus précisément, la segmentation sémantique.

Bien que ce modèle offre des avantages intéressants, il présente également certaines limitations, notamment en ce qui concerne son évolutivité, un aspect crucial pour des applications à plus grande échelle. Dans ce cas, en traitant des images de grandes tailles, on sera face à un coût de calcul élevé, ce qui implique la difficulté de capturer les informations plus fines. Le ViT divise l'image d'entrée en plusieurs régions de tailles fixes, facilitant ainsi l'extraction des caractéristiques locales. Ces derniers sont égaux à 16x16 pixels. Les auteurs de l'article ont fixé cette valeur puisqu'elle a abouti à de bonnes performances sur un ensemble de données de référence qui est ImageNet.

Après la division de l'image, ces segments seront introduits dans un réseau de neurones appelé Transformer. Cette approche pourra être appliquée pour des données de petite taille, mais si on se retrouve face à un grand ensemble de données, la tâche deviendra beaucoup plus difficile, vu qu'on sera face à un coût de calcul aussi élevé. Si on prend à titre d'exemple une image de taille deux fois supérieure à celle de référence, dans ce cas le nombre de régions sera lui-même multiplié par 4. On pourra donc conclure que l'architecture de ViT est beaucoup plus prohibitive pour les images de grande taille. Dans ce cas, notre modèle aura des difficultés pour capturer les informations spatiales entre les parcelles, donc un

impact négatif sur sa capacité à classer l'image avec précision.

Pour conclure, une des plus grandes limitations du ViT est son évolutivité face aux images plus grandes. Ce qui pourrait créer un obstacle dans certains domaines d'application, comme dans le domaine médical ou l'imagerie satellitaire où les résolutions de l'image sont beaucoup plus importantes. Afin de faire face à cette contrainte, une étude a proposé une méthode innovante qui emploie un modèle hiérarchique d'un réseau capable de traiter des images de grande taille tout en préservant une grande précision. Le modèle Swin Transformer renforce l'efficacité tout en réduisant les limitations liées à l'analyse d'images de haute qualité (Liu *et al.* (2021b)).

1.9.1 Mode fonctionnement de Swin Transformer

Les images à haute résolution sont traitées à différents niveaux d'abstraction selon une série d'étapes dans cette architecture, comme c'est indiqué dans la figure 1.10 :

a. **Division en plusieurs patchs :**

   L'image d'entrée RGB sera divisée en un nombre fixe de régions qui ne se chevauchent pas à l'aide d'un module de division similaire à celui de ViT. La taille de la région est un hyperparamètre qui peut être ajusté en fonction des dimensions de l'image d'entrée et des spécificités de la tâche à accomplir. En général, on opte pour une région de taille 4×4, dont la dimension totale est égale à $4 \times 4 \times 3 = 48$. Ainsi, chaque patch est considéré comme un 'token' et incorporé dans le Swin Transformer pour un traitement ultérieur.

b. **Linear embedding :**

   Dans le cadre de l'architecture du Swin Transformer, chaque région, représentée par des tableaux 2D de valeurs de pixels, sera transformée en un vecteur. Cependant, ces vecteurs 1D ne peuvent pas être directement introduits dans l'architecture Transformer. Et donc, ils seront projetés dans un espace d'intégration à haute dimension à l'aide d'une projection linéaire pouvant être apprise. Le but de cette étape est de faire apprendre à Swin Transformer les informations les plus significatives et les plus abstraites de l'image d'entrée, d'où une meilleure classification avec précision. Il faudra

noter à ce propos que la matrice de projection est un paramètre du modèle appris tout au long de la phase d'entraînement.

c. **Shifted Window Attention :**

On pourra clairement voir dans l'architecture de Swin Transformer qu'il contient plusieurs blocs. À chaque bloc, les régions voisines seront regroupées en un ensemble de fenêtres qui se chevauchent. Pour produire une représentation hiérarchique, à chaque stade, le nombre d'éléments est réduit en les concaténant au fur et à mesure que le réseau devient plus profond. Le nombre de régions est ainsi réduit d'un multiple de $2\times2 = 4$.

La taille de chaque fenêtre est également un hyperparamètre qui peut être réglé. Par la suite, Swin Transformer va appliquer un mécanisme d'auto-attention à chaque fenêtre pour calculer un ensemble de caractéristiques au niveau de la fenêtre. Le but de cette étape est de garantir que toutes les informations des fenêtres voisines sont immergées dans le calcul.

d. **Transformer Encoder :**

Le bloc Swin Transformer se présente comme étant un encodeur multicouche qui va traiter les caractéristiques présentes au niveau de notre fenêtre de dimension 1D pour capturer les informations les plus précises se retrouvant dans notre image d'entrée. Le codeur Transformer se compose de plusieurs blocs, chacun contient un mécanisme d'attention et un réseau neuronal de type feedforward. La sortie de chaque bloc est ajoutée à l'entrée et le résultat passe par une opération de normalisation de la couche.

e. **Down-Sampling :**

Après le traitement de chaque étape, la résolution des caractéristiques au niveau de la fenêtre est réduite par un facteur de deux en appliquant une opération de mise en commun maximale sur les caractéristiques au niveau de la fenêtre. Cela a pour effet de comprimer les informations contenues dans les cartes de caractéristiques. Cela permet de réduire le coût de calcul du traitement des images à haute résolution et de capturer des informations contextuelles de plus en plus globales.

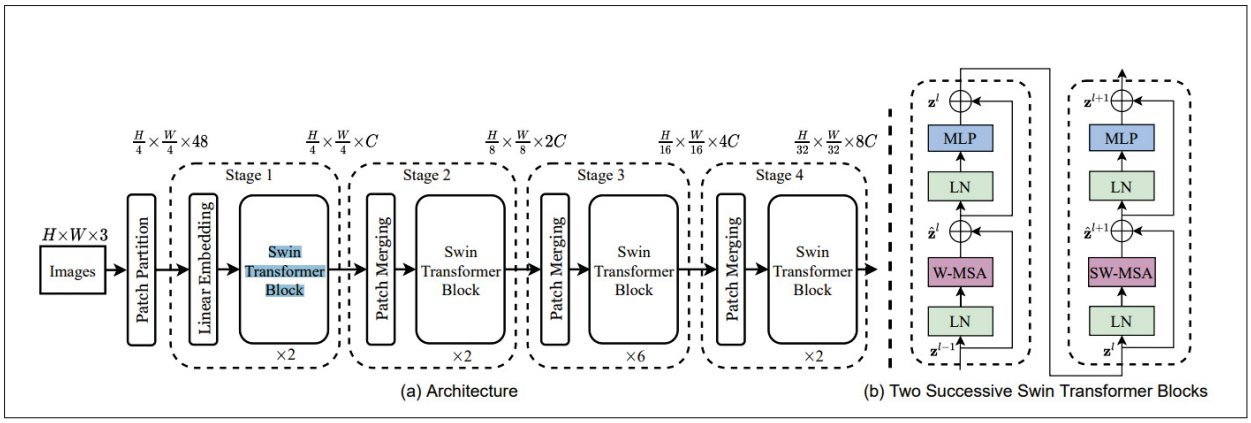

Figure 1.10 Architecture du modèle Swin Transformer
Tirée de l'article Swin Transformer: Hierarchical Vision Transformer using Shifted Windows

1.10 Analyse comparative

Dans cette partie, nous examinerons de manière comparative les divers modèles présentés, en soulignant leurs similitudes et leurs disparités. Ils sont basés sur une architecture Transformer et utilisent des mécanismes d'auto-attention pour traiter les images. L'objectif de cette comparaison est de mettre en lumière les points forts et les limitations de chaque modèle, ainsi que leur pertinence respective dans le domaine des applications d'imagerie médicale. Il est également important de souligner que Swin Transformer favorise la fusion hiérarchique des caractéristiques, tandis que SegFormer opte pour collaborer avec les pyramides spatiales.

D'autre part, ViT se focalise sur la combinaison de la moyenne globale afin d'intégrer les caractéristiques. Swin Transformer déploie une auto-attention basée sur les fenêtres avec des tailles de régions plus petites et convient aux ensembles de données plus importants. Tandis que ViT et SegFormer déploient une auto-attention à plusieurs têtes avec des tailles de régions plus grandes et conviennent aux ensembles de données plus petits. Enfin, Swin Transformer utilise l'intégration de la position absolue, par opposition à l'intégration de la position apprenable utilisée par ViT et SegFormer.

Tableau 1.1 Comparaison entre les différents modèles

Nom du modèle	ViT	SegFormer	SwinTransformer
Architecture	Architecture Transformer	Architecture Transformer	Architecture Transformer
Mécanisme	Multi-head self-attention	Multi-head self-attention	Window-based self-attention
Taille de l'image	Variable	Variable	Fixe
Taille de la région	16x16 ou 32x32	16x16 ou 32x32	4x4
Compléxité	Faible	Faible	Plus élevée
Performance	Exactitude plus faible	Exactitude plus faible	Exactitude plus élevée
Application	Petite base de données	Petite base de données	Large base de données

1.11 Approches combinées de CNN et ViT pour l'analyse d'images

Certaines études se sont inspirées de l'architecture de ViT afin de les combiner avec un décodeur pour l'adapter à des problèmes spécifiques. Une des approches récentes consistait à prendre le Swin Transformer avec Knet comme étant une alternative pour la segmentation ventriculaire gauche en échocardiographie. Cette combinaison traite les images de fin de diastole et de systole afin de calculer la fraction d'éjection. L'architecture de ce modèle repose principalement sur deux modules principaux pour former un encodeur-décodeur. Le bloc Swin transformer est efficace pour extraire les caractéristiques hiérarchiques à multi-échelles tout en réduisant la complexité computationnelle. K-Net aura recours à une tête de décodeur itérative pour fournir le résultat de la segmentation sémantique du ventricule gauche de manière progressive et précise. Cette association a permis de configurer une méthode alternative de pur ViT afin de profiter de sa capacité à saisir les relations spatiales globales, tout en offrant des résultats plus précis grâce au mécanisme itératif du décodeur (Liao *et al.* (2023)). Bien que la tâche de segmentation ou de détection constitue un élément fondamental pour l'analyse des images, il

serait également intéressant d'incorporer des informations spatio-temporelles pour améliorer l'évaluation ventriculaire. C'est ainsi que les ViT démontrent leur efficacité pour le suivi temporel, permettant une évaluation continue et précise des indices cliniques au fil du temps. En effet, les méthodes basées sur le suivi à partir d'une seule image ont prouvé leur efficacité dans certaines applications, mais elles présentent des limitations importantes en termes de précision. On manque souvent de cohérence temporelle, où la position de notre objet d'intérêt change de manière régulière, ce qui introduira des erreurs significatives (Taskén *et al.* (2024)). Une étude exploite ainsi l'efficacité de ViT pour un suivi efficace, ajoutant ainsi une dimension dynamique à l'évaluation des images. C'est dans ce cadre que l'architecture MixFormer se présente comme étant une structure de suivi simple conçue pour unifier l'extraction des caractéristiques et l'intégration de la cible à l'aide d'une architecture basée sur un ViT. L'algorithme de suivi est initialisé par une seule image qui identifie le point de départ. Cette position permettra au modèle de faire le suivi tout au long de la séquence. Grâce à son élément clé, le module d'attention mixte (MAM), MixFormer assure une localisation précise et continue des objets (Cui, Jiang, Wu & Wang (2024)). D'où un modèle robuste, notamment dans des environnements cliniques complexes comme le suivi de l'anneau mitral en échocardiographie. De même, afin d'acquérir une vision complète et compréhensive de l'image, certaines recherches se sont tournées vers une approche innovante pour exploiter les capacités des CNN et ViT. Le module SMA (Self-adaptive Multi-Attention) offre justement cette possibilité, en introduisant de manière adaptative les mécanismes d'attention basés sur les CNN et les ViT. Le module SMA a été développé dans le but d'assurer une opération de fusion et de détection efficace des cartes caractéristiques. Le modèle CardSegNet a pu améliorer sa capacité à détecter les contours et les structures complexes du cœur grâce à son mécanisme. En incorporant cette fonctionnalité de fusion adaptative, l'architecture CardSegNet améliore la compréhension des caractéristiques à diverses échelles, ce qui se traduit par une segmentation plus précise et solide. Cela a été développé et évalué à partir de deux ensembles de données publics, à savoir ACDC 2017 et MMs-2. De cette manière, cette solution prometteuse nous offre une segmentation automatique, rapide et fiable, ce qui est crucial pour le diagnostic des maladies cardiaques (Aghapanah *et al.* (2024)).

1.12 Conclusion

Dans le cadre de cette revue de littérature, on a essayé de mettre l'accent sur l'importance de l'échocardiographie, en mettant l'accent sur l'étude du cœur. L'étude de cet organe vital est nécessaire pour saisir son fonctionnement complexe et mettre en lumière les relations entre les différentes chambres. En tant qu'outil d'imagerie, l'échocardiographie joue un rôle essentiel dans l'évaluation de la fonction et de la morphologie du cœur, en fournissant des informations cruciales pour le diagnostic et le suivi des affections cardiaques. À cet égard, on a essayé d'examiner certaines études de recherche ayant traité la tâche de la segmentation du ventricule gauche à partir des images échocardiographiques. Les architectures CNN ont prouvé des résultats prometteurs, en particulier en ce qui concerne la détection de formes et de motifs locaux dans les images médicales. Toutefois, ils ont également révélé certaines limitations en ce qui concerne l'adaptation et l'efficacité des modèles face à la complexité des données échocardiographiques. Face à ces limitations, on a fait introduire les modèles ViT qui ont prouvé des progrès notables dans l'amélioration de la précision de segmentation en ce qui concerne leur capacité à capturer des relations à longue portée dans les images, offrant ainsi une meilleure compréhension contextuelle. C'est dans cette perspective que notre travail se présente en combinant les avantages des CNN et la capacité des ViT. On aura recours à une architecture novatrice qui répond aux défis posés par l'analyse des images échocardiographiques.

CHAPITRE 2

AUTOMATIC SEGMENTATION OF ECHOCARDIOGRAPHIC IMAGES USING A SHIFTED WINDOWS VISION TRANSFORMER ARCHITECTURE

[1] Souha NEMRI , [2] Luc DUONG

[1,2] Département de génie logiciel et des TI, École de technologie supérieure,
1100 Notre-Dame Ouest, Montréal, Canada H3C 1K3
Article publié dans la revue « Biomedical Physics, Engineering Express », Septembre 2024

2.1 Abstract

Echocardiography is one the most commonly used imaging modalities for the diagnosis of congenital heart disease. Echocardiographic image analysis is crucial to obtaining accurate cardiac anatomy information. Semantic segmentation models can be used to precisely delimit the borders of the left ventricle, and allow an accurate and automatic identification of the region of interest, which can be extremely useful for cardiologists. In the field of computer vision, convolutional neural network (CNN) architectures remain dominant. Existing CNN approaches have proved highly efficient for the segmentation of various medical images over the past decade. However, these solutions usually struggle to capture long-range dependencies, especially when it comes to images with objects of different scales and complex structures. In this study, we present an efficient method for semantic segmentation of echocardiographic images that overcomes these challenges by leveraging the self-attention mechanism of the Transformer architecture. The proposed solution extracts long-range dependencies and efficiently processes objects at different scales, improving performance in a variety of tasks. We introduce Shifted Windows Transformer models (Swin Transformers), which encode both the content of anatomical structures and the relationship between them. Our solution combines the Swin Transformer and U-Net architectures, producing a U-shaped variant. The validation of the proposed method is performed with the EchoNet-Dynamic dataset used to train our model. The results show an accuracy of 0.97, a Dice coefficient of 0.87, and an Intersection over Union (IoU) of 0.78. Swin Transformer models are promising for semantically segmenting echocardiographic images and may help assist

cardiologists in automatically analyzing and measuring complex echocardiographic images.

Keywords : Echocardiography, Semantic segmentation, Left Ventricle, Transformers, U-Net

2.2 Introduction

Congenital Heart Disease (CHD) is the most common type of birth defect among humans, occurring in 0.5-0.8% of all live births, and affecting 1.5 million children worldwide (Hoffman & Kaplan (2002))(Reller, Strickland, Riehle-Colarusso, Mahle & Correa (2008)). CHD prevalence is estimated to be 8 cases per 10,000 live births in the population. A major challenge when diagnosing a complex CHD is visualizing anatomical structures.

Echocardiography is an imaging method generally used to acquire anatomical data from the heart. It is a very simple technology used by cardiologists to visualize the heart's 4 chambers. It provides a representation of the heart's movements, producing images of the heart's valves and chambers, without the need for radiation. It allows the cardiologist to visualize the heart and assess its contraction and relaxation, as well as the valve function. The type of echocardiography the patient undergoes may vary as a function of the information needed by the clinician. The left ventricular volume and ejection fraction are two essential volumetric analyses that provide a detailed understanding of cardiac contractility, which leads to better cardiac function diagnosis. Echocardiographic image segmentation can play a significant role in the automatic analysis and diagnosis of cardiac function. Precise segmentation of anatomical structures in medical images is an essential task for the clinical treatment of certain cardiac diseases. Some cardiac parameters such as the volumes of end systolic and end diastolic, ejection fraction and myocardium mass are good indicators of cardiac health, representing reliable diagnostic value. Clinicians can benefit from the advantages of segmentation to calculate these clinical measurements which are essential for any surgical intervention and treatment follow-up. Recent studies have shown segmentation to be essential and useful for extracting anatomical structures, facilitating the study of medical phenomena and the discovery of new treatments. In most clinical settings, the cardiologist or a trained operator still performs the segmentation step manually, which is

laborious and time-consuming, as well as being subject to inter- and intra-observer variability. So automatic segmentation would help physicians in their decision making.

Some research efforts have focused on deep learning to study left ventricular segmentation and to calculate clinical measures for heart disease diagnosis. In particular, the Multi-attention Efficient Feature Fusion Network has been used for automatic segmentation in echocardiography. It incorporates a deep supervision mechanism and spatial pyramid feature fusion to improve feature extraction (Zeng *et al.* (2023)). Similarly, the calculation of the left ventricular volume (LVEF) represents an effective measure for assessing cardiac health in children, with the deep learning model being adapted to pediatric data. In this context, physiological variations in children are taken into account, and consequently, the model provides an acceptable clinical error and supports the independent assessment of LVEF (Zuercher *et al.* (2022)). Similarly, various projects have been proposed using the U-Net model and its variants, and have achieved good results. The DPS-Net algorithm, based on the U-Net architecture, has shown effectiveness in left ventricle segmentation and ejection fraction measurement across different heart disease phenotypes (Liu *et al.* (2021a)). However, these methods often ²require a large number of ground-truth labels, which is time-consuming. To address this, researchers proposed a method that combines multi-level and multitype self-generated knowledge, using a superpixel approach and various pretext tasks (Yu *et al.* (2023)).

Convolutional neural network (CNN) models have been the most commonly used models in many applications (Aubry & Duong (2023)). In the field of computer vision, CNN architectures remain dominant. They have become the cornerstone of a lot of tasks due to their ability to learn the most important features from our input data. Existing CNN approaches have even proved highly efficient for the segmentation of various medical images over the past decade. Recently, Vision Transformers (ViT) have seen their interest grow significantly in the field of computer vision (Dosovitskiy *et al.* (2021)). The reliance on CNN is not even necessary and a pure transformer applied directly to sequences of image patches can perform very well on images. CNN usually struggle to capture long-range dependencies, especially when it comes to images with objects of different scales and complex structures.

The basic ViT model takes the input image and divides it into several fixed size patches, which are inputted into a neural network. This task is straightforward for small images, but can be computationally intensive for larger ones, such as medical images. With the latter, the basic ViT might fail to capture the spatial information between patches, which may have a negative impact on its ability to accurately perform segmentation. To tackle this limitation, the Shifted Windows (Swin) is proposed to better handle the segmentation of larger images while maintaining a high accuracy. Its hierarchical architecture, which is based on the Swin concept (Liu *et al.* (2021b)), processes high-resolution images by analyzing them in a series of stages at different levels of abstraction. The Swin Transformer's architecture is founded on the Shifted Window Attention (SWA) concept. It groups neighboring patches into a set of overlapping windows. In this context, the algorithm does not apply attention mechanism in a standard fashion ; rather, the SWA allows each patch to focus more on patches that are spatially close to it. Patches that resemble each other or are spatially close to one another will have stronger relationships thanks to SWA. These stronger relationships are then used in the U-Net part to perform semantic segmentation.

The SWA of the Swin Transformer thus ensures a better detection of local relationships between patches, which is beneficial for the image segmentation task. The features extracted by the Swin Transformer can then be used by the U-Net part of the Swin U-Net to perform semantic segmentation. This method delivers excellent results in complex image segmentation tasks.

The goal of this study is to evaluate a Swin U-Net architecture for the segmentation of echocardiographic images. The study is organized as follows : first, the methodology and the datasets used to validate the proposed approach are described, followed by a presentation of the evaluation of the model. Finally, a discussion and conclusion is provided.

2.3 Methodology

2.3.1 Database

Two public datasets were used to evaluate the performance of our proposed approach. The first was the Echonet Dynamic dataset (Ouyang *et al.* (2020)), a database designed specifically for the interpretation and analysis of echocardiographic images. Its information allows to visualize the structure of the human heart, while evaluating its function. It contains around 10,030 apical four-chamber echocardiography videos. This dataset was collected at Stanford University Hospital between 2016 and 2018 from medical examinations performed on patients. Each video features a four-chamber apical view of the heart, as shown in Figure 2.1 below.

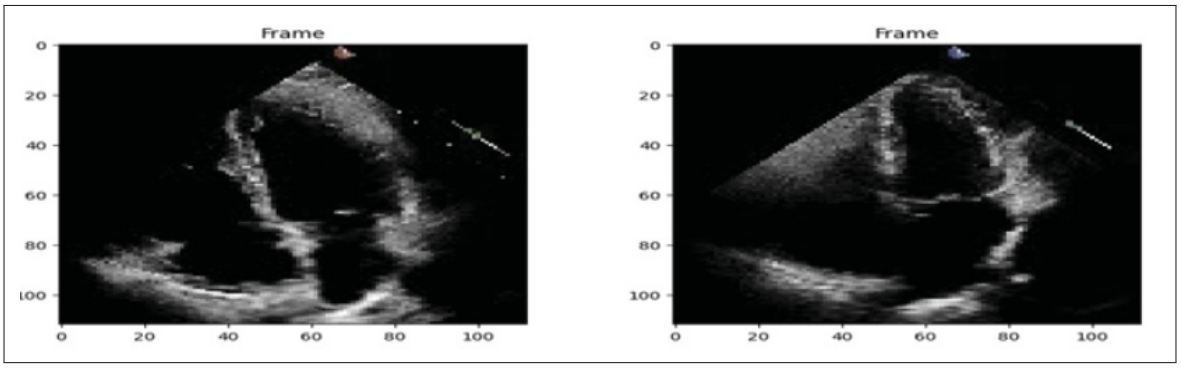

Figure 2.1 Sample of our database (EchoNet-Dynamic)

For each video sequence, a wide range of information is available, which is useful for diagnostics and follow-up of different cardiac diseases. For example :

1. **Number of frames :** This indicates the total number of frames in each video.
2. **Split :** Assigning video to Train or Valid or Test datasets
3. **Frame :** Frame number on which left ventricular segmentation tracing was performed

Next, we used the EchoNet-Pediatric dataset (Reddy, Lopez, Ouyang, Zou & He (2023a)). It consists of a set of echocardiogram videos labeled by human experts such as to give us idea of the assessment of left ventricular function. It was obtained at Lucile Packard Children's Hospital

Stanford in the context of routine clinical care, from 2014 to 2021, and from children aged between 0 and 18 years of age, and of different sizes. It contains two-dimensional grayscale clips of A4C (apical four-chamber) and PSAX (parasternal short-axis) views. For our case, however, we focus only on the four-chamber view in order to perform a comparative analysis with the results obtained with the EchoNet-Dynamic dataset. This dataset contains both anatomically normal hearts with normal ejection fraction and patients.

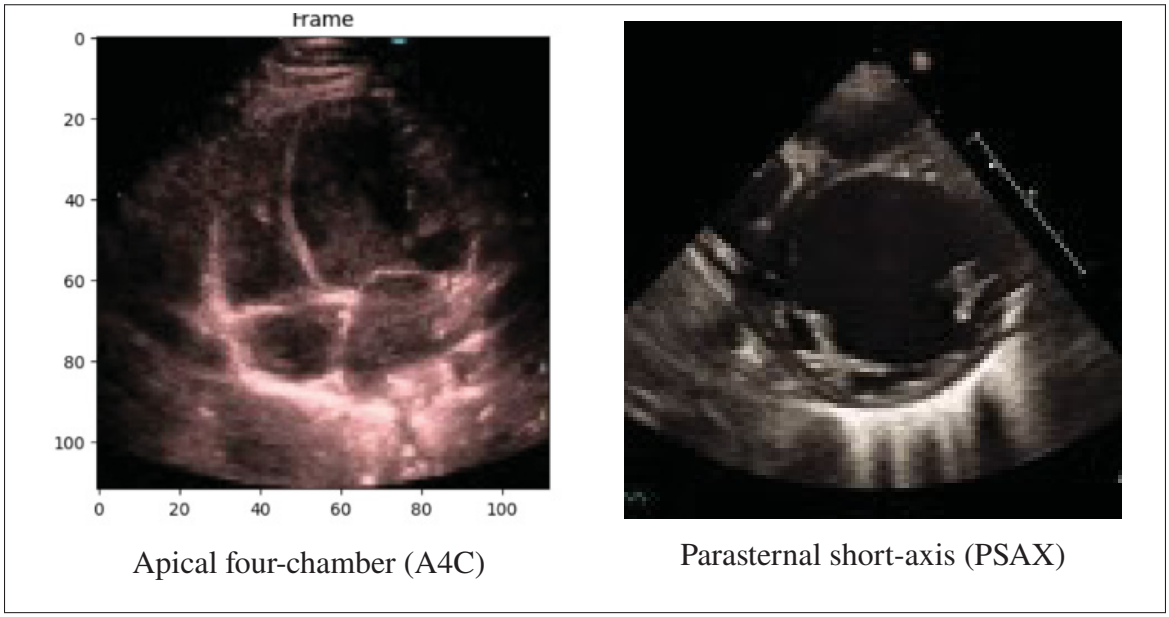

Figure 2.2 Sample of our database (EchoNet-Pediatric)

For both databases, the image sequence consists of a series of 112 by 112 pixel grayscale images. In order to delimit the boundary of the left ventricular cavity wall, it was essential to generate the ground truth masks for our images. To this end, we used the coordinates X1, Y1, X2 and Y2 of the EchoNet-Dynamic dataset.

These coordinates are connecting the most distant points on the ventricle surface to create our line segment, whereas for EchoNet-Pediatric, we used the X and Y columns to generate the corresponding masks. Overlaying the original image with our ground truth mask allows to see the visual results, as illustrated in Figures 2.3 and 2.4. We have further displayed the contours of the segmentation mask to better visualize the results. This step is intended to validate that the mask

we generated corresponds perfectly to the original image and that the predicted segmentation after training also corresponds to the original images.

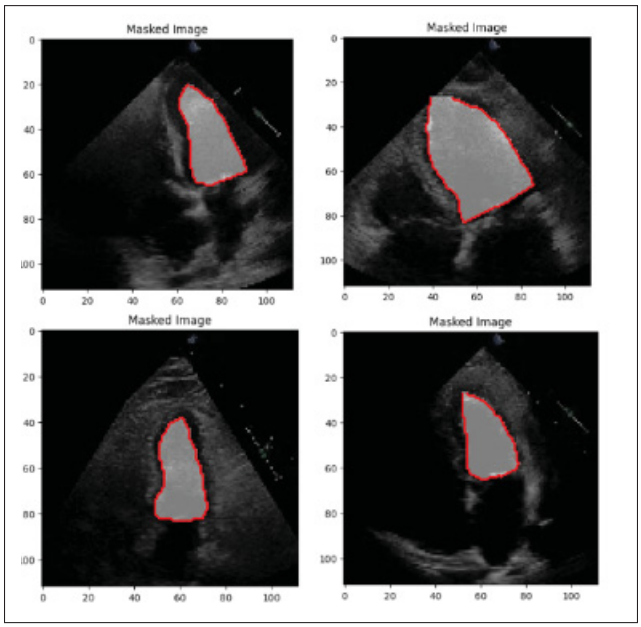

Figure 2.3 Masked images of EchoNet-Dynamic Dataset

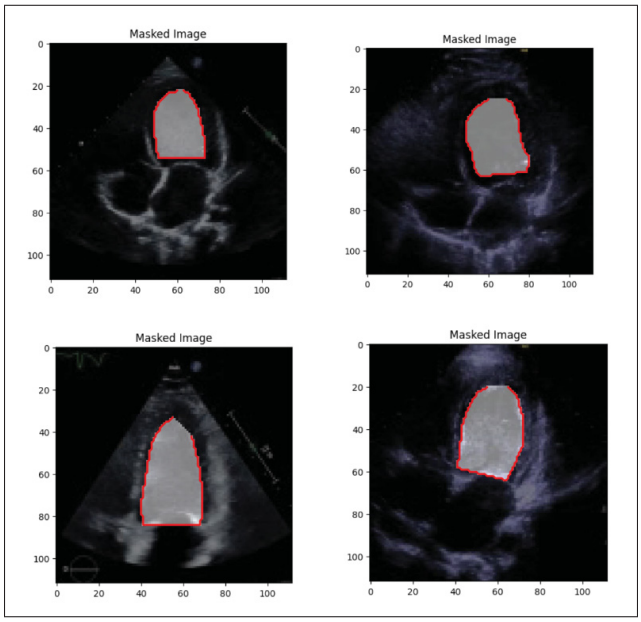

Figure 2.4 Masked images of EchoNet-Pediatric Dataset

2.3.2 Preprocessing

For each of our databases, there is a total of 15,000 images for the Echonet Dynamic dataset and 6415 for the Echonet Pediatrics dataset. For each of these, we opted to split the data as follows : 70% for training our model, 20% for validation and 10% for testing. All our data loaders were resized to match the input of our Swin Unet model. Our data's dimensions were (256,256,3), while the size of the masks was set fixed at (256,256,1). We then proceeded to normalize our entire database. This step is crucial for any deep learning problem. In this case, all our images were on a similar scale, standing between 0 and 1. Normalization was necessary to allow the neural networks to more easily ensure that the features were on a similar scale. This would guarantee the stabilization of the descent gradient and allow the optimization algorithm to converge much faster and more reliably, avoiding oscillation and slow convergence due to non-normalized or differently scaled features. We used the **Min-max normalization (feature scaling)** method, which performs a linear transformation on the original data. This is a common technique used to transform data into a range generally between 0 and 1. This transformation proceeds through the following formula :

$$NormalizedImage = \frac{ResizedImage - min(ResizedImage)}{max(ResizedImage - min(ResizedImage))} \tag{2.1}$$

Following this operation, we calculated the minimum value of our resized image, as well as its maximum value. The result was then subtracted and divided by the data range, giving values between 0 and 1. Our images and masks were then transformed to a common range, simplifying the learning process for the models to be trained. This ensured that features with high values would not dominate during the training process, thus improving the model stability.

2.3.3 Swin U-Net Architecture

A new deep learning model concept is proposed for the segmentation of echocardiographic images, through the Swin U-Net algorithm, which combines the architecture of a CNN encoder-decoder with the structure of a transformer (Cao *et al.* (2021)). This U-shaped variant combines

the advantages of the Swin Transformer and of U-Net architectures for semantic segmentation tasks. It allows to capture long-term dependencies using the Swin Transformer's self-attention mechanism, while preserving the high-resolution feature representation offered by U-Net.
The original image is initially divided into a set of patches, each of which is processed independently in the transform blocks. Thanks to the auto-attention mechanism, these blocks capture the local relationships within each patch. First, the Swin Transformer encoder processes the input image. Its role is to capture global dependencies and extract high-level features and transmit them to the decoder. Once the decoder has extracted the spatial details, it generates the final segmentation mask. **The encoder** processes the input image in a series of convolution and pooling operations, progressively reducing the spatial resolution. This process allows the model to capture patterns and more complex semantic information, as well as to extract abstract features. After training, we obtain a rich representation of image features. These feature maps have a much smaller spatial dimension and are transmitted to the decoder part of Swin U-Net, where spatial detail is restored.

These low-resolution feature maps, which have been generated at various stages, are now merged together. This process combines information from different scales, thus improving the model's ability to capture both global and local contexts. This fusion enables Swin U-Net's architecture to efficiently capture both global contextual information and the finest details. This is known as the **Patch Merging Layer** step.

Afterwards, Swin U-Net employs the U-Net structure's **decoder**, which performs upsampling operations. The aim is to increase the spatial resolution of the feature maps received by the encoder. Also, the presence of **the skip connections** in our model is essential, as they link the encoder to the corresponding decoder layers. These connections are responsible for merging the multi-scale features from the encoder with the upsampled features. This action is necessary in order to reduce the loss of spatial information caused by the downsampling in the first part. With these connections, the decoder can simultaneously combine high-level, context-rich information from the encoder with fine, spatially-detailed information from previous layers. This allows the

network to carry out more precise predictions and accurately capture complex details in the output. Ultimately, we will be able to recover the finest spatial details that have been lost.

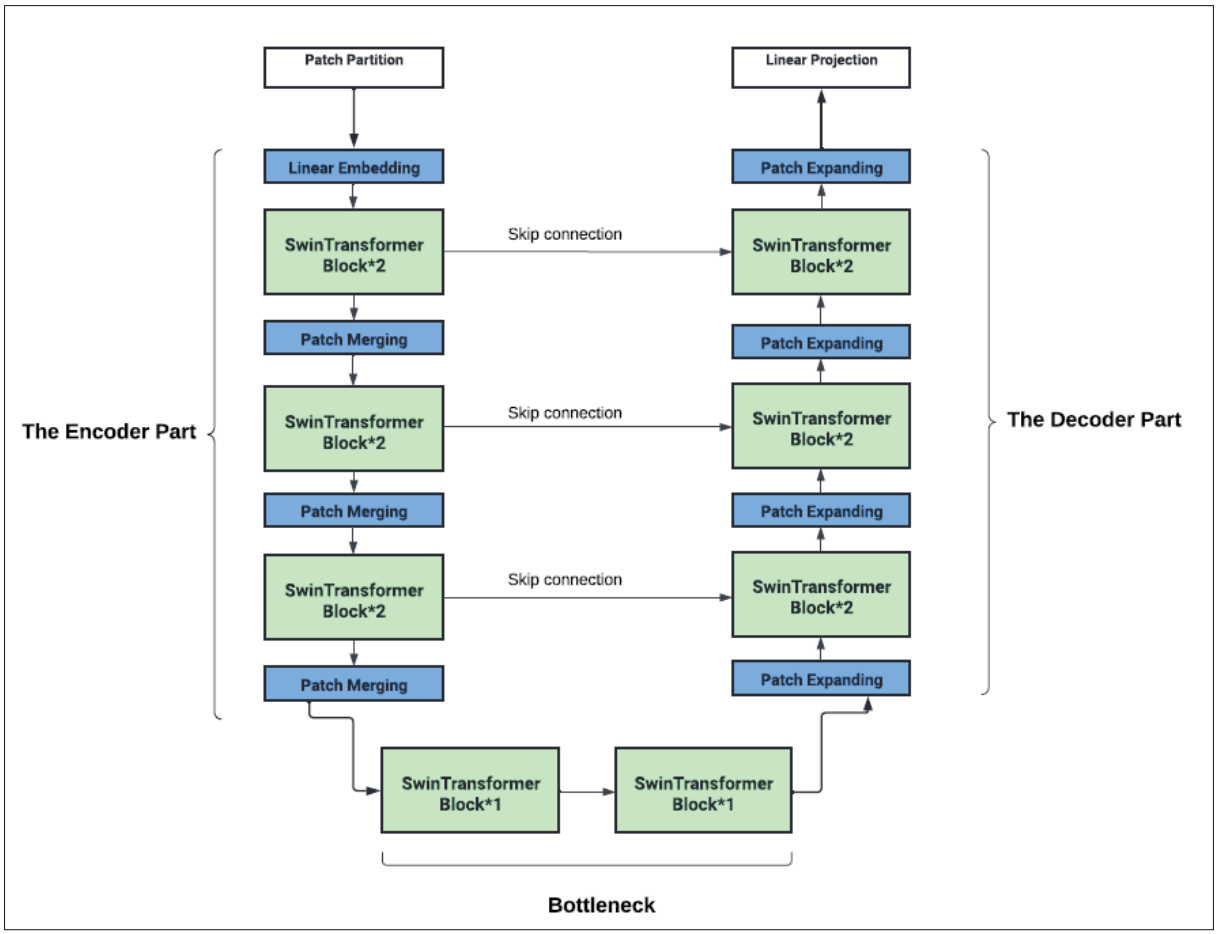

Figure 2.5 Swin U-Net Architecture

Finally, we add the **Patch Expanding Layer**, which will restore the resolution of the feature maps to the input level. The Linear Projection Layer, for its part, will produce segmentation predictions at the pixel level.

From Figure 2.6, it can be seen that **the Swin Transformer block** is the basic unit of a symmetric Encoder-Decoder architecture with skip connections, named Swin U-Net. We note the presence of two consecutive Swin Transformer blocks, each composed of a LayerNorm (LN) layer, a multi-headed self-attention module, a residual connection and a two-layer MLP with GELU non-linearity. The two transformer blocks that follow implement the multihead window-based

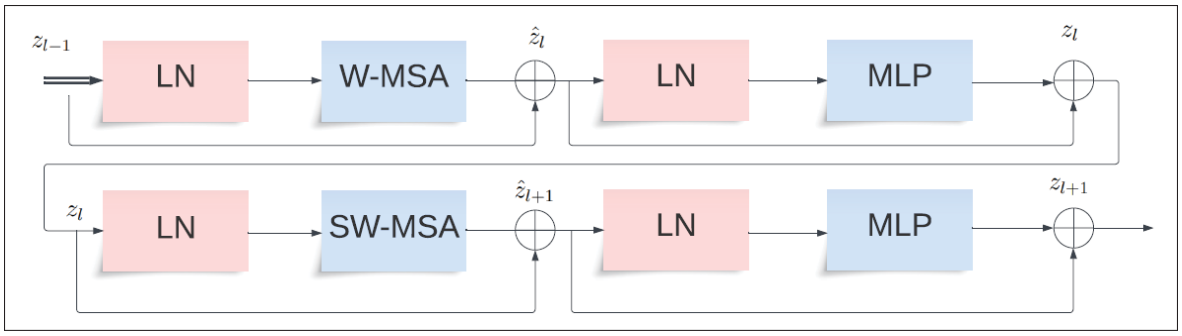

Figure 2.6 Swin transformer block

self-attention module (W-MSA) and the multihead offset window-based self-attention module (SW- MSA), respectively :

$$\hat{z}_l = \text{W-MSA}(\text{LN}(z_{l-1})) + z_{l-1} \tag{2.2}$$

$$z_l = \text{MLP}(\text{LN}(\hat{z}_l)) + \hat{z}_l \tag{2.3}$$

$$\hat{z}_{l+1} = \text{SW-MSA}(\text{LN}(z_l)) + z_l \tag{2.4}$$

$$z_{l+1} = \text{MLP}(\text{LN}(\hat{z}_{l+1})) + \hat{z}_{l+1} \tag{2.5}$$

Given that $\hat{z}_l$ is the output of W-MSA and SW-MSA, while z_l is the result of MLP module.

2.3.4 Implementation details

To define our architecture and make it suitable for the segmentation task, we opted for a value of 1 for the variable n-labels, since this is a binary segmentation and our model will predict a single mask. Our patch size used by Swin Transformers was set to (4, 4), meaning that the input image would be divided into 4x4 patches for initial processing. Prior to the start of our training, various loss functions commonly used for segmentation tasks were available, allowing to compile the model with an appropriate loss function. We chose **Binary cross entropy**, which is commonly used for binary segmentation. In our case, the goal was to delineate the left ventricle of the human heart. This option is robust to imbalance, as the background pixels far outnumber the

foreground pixels. **BCE** manages this imbalance reasonably well and penalizes false positives and false negatives effectively, thus promoting balanced segmentation. To be able to evaluate the performance of our model at the end of the training, we had to define common metrics for semantic segmentation throughout the process. **Dice Coefficient (DSC)** or F1 is the score most widely used to measure model performance for the medical image segmentation task. Our aim was to observe the overlap between the predicted segmentation and the ground truth. In other words, we can simplify DSC to the following equation :

$$DiceCoefficient = \frac{2*AreaofOverlap}{TotalArea} \tag{2.6}$$

We also used the IoU function, also known as the Jaccard index, which is another evaluation measure commonly used in segmentation tasks :

$$IoU = \frac{AreaofOverlap}{TotalOfUnion} \tag{2.7}$$

We have incorporated the Hausdorff distance analysis in order to provide a more comprehensive assessment of segmentation accuracy by quantifying the maximum discrepancy between two sets of points. In this case, the sets A and B are respectively the ground truth and the predicted values.

$$H(A,B) = \max\left\{\sup_{a\in A}\inf_{b\in B} d(a,b), \sup_{b\in B}\inf_{a\in A} d(a,b)\right\} \tag{2.8}$$

2.4 Results

In Figure 2.7, it can be seen that Swin U-Net has produced a predicted mask that approximates the ground truth segmentation mask. From a visual standpoint, there is a good overlap between the results of our model and the images in the test database. This points to a good delimitation of the left ventricular border achieved by Swin U-Net.

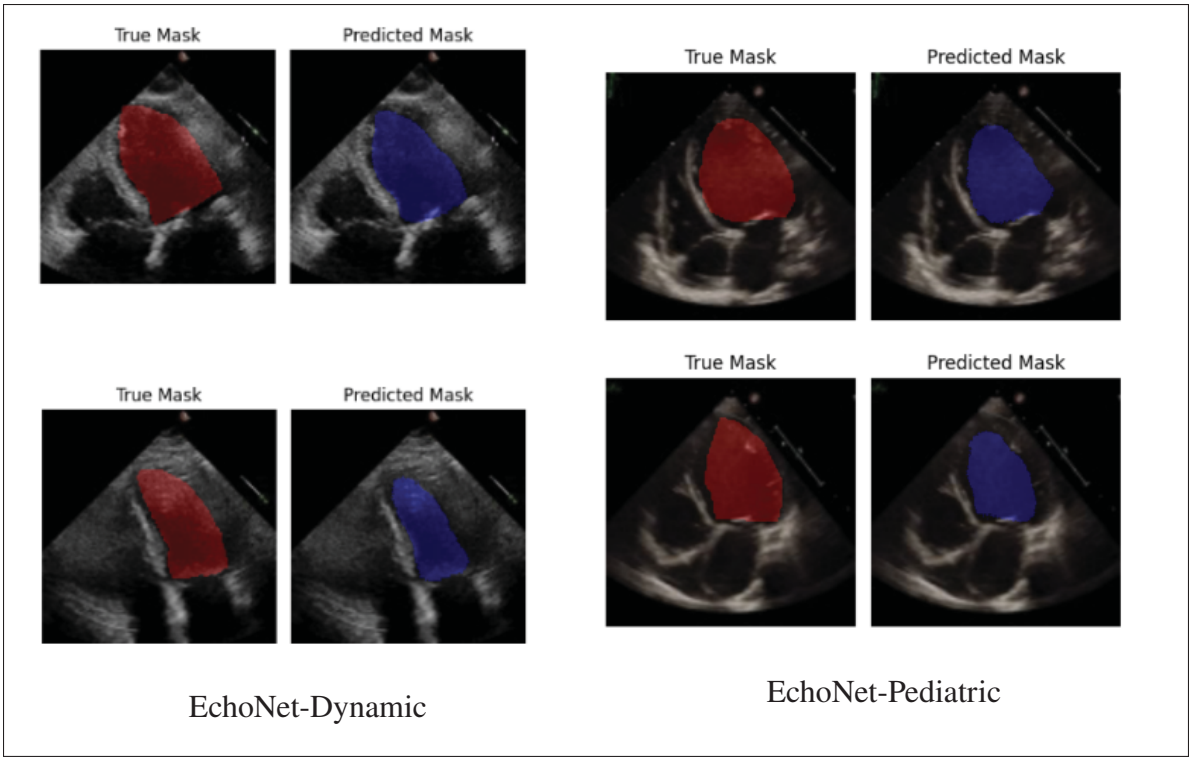

Figure 2.7 Representation of the segmentation of the left ventricle with Swin U-Net (blue) and its ground truth (red)

During the training phase, IoU and DSC were our basic metrics. Our index provides a perspective on the quality of segmentation predictions and complements the Dice coefficient as a valuable metric for monitoring the performance of our model. Figure 2.8 shows the results obtained by the scores at the end of the training phase. A progressive decrease can be seen in the loss function, leading to a reduction in the error between predictions and actual values. Towards the end, we were able to reach a value equal to 0.04. On the other hand, we cannot settle for the loss function reduction alone, as it is not sufficient to guarantee a high-performance model.

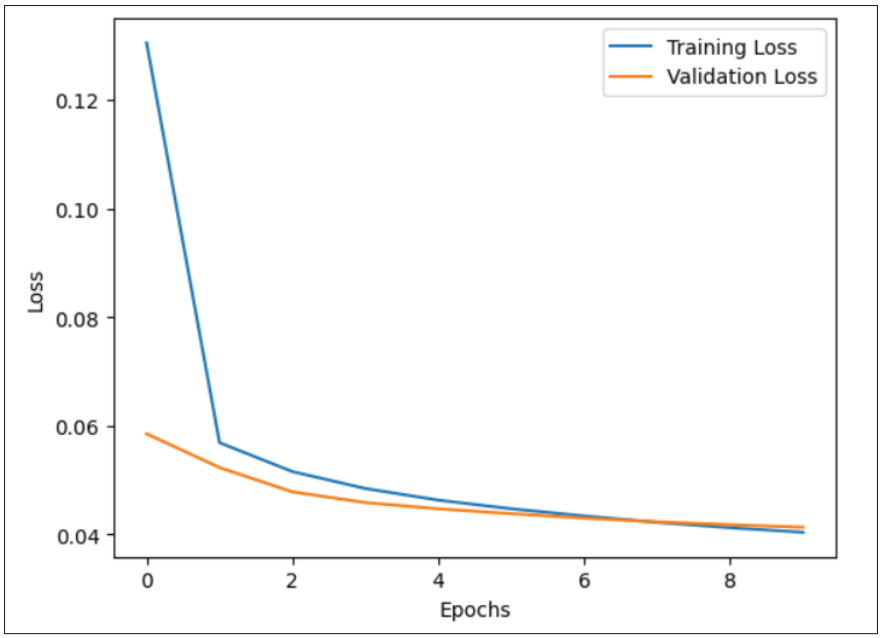

Figure 2.8 Swin-Unet results (EchoNet-Dynamic)

The IoU and Dice coefficient scores continue to increase progressively. This indicates a potential overlap between the predicted and the true masks. Our Swin Unet model accurately segments the left ventricle with a Dice coefficient of 0.88 and an IoU score of 0.78.

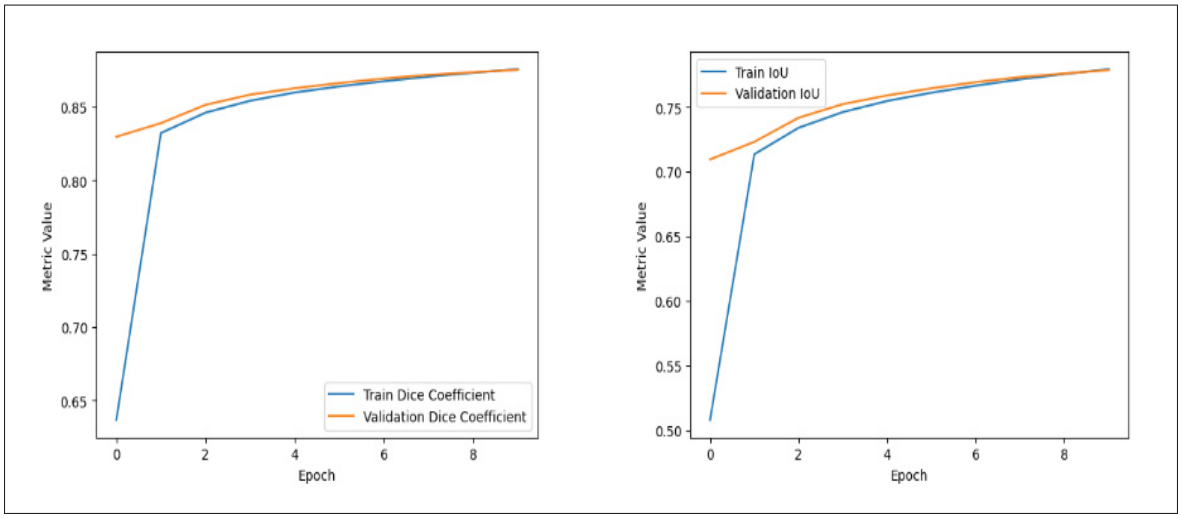

Figure 2.9 Swin-Unet results (EchoNet-Dynamic)

We have repeated the same process for our second dataset Echonet Pediatric in order to perform a comparative analysis with the results obtained with the EchoNet-Dynamic dataset. Figure 2.10 illustrates the results of segmentation of our model Swin Unet which indicates the effectiveness of our model in pediatric echocardiography. The model achieved a Dice coefficient of 80.94% which proves a good overlap between the prediction of our model and the ground truth mask. Additionally for our metric Intersection over unit (IoU), with a score approximately equal to 70%, it further confirms the ability of our model to perform a good segmentation of the left ventricle.

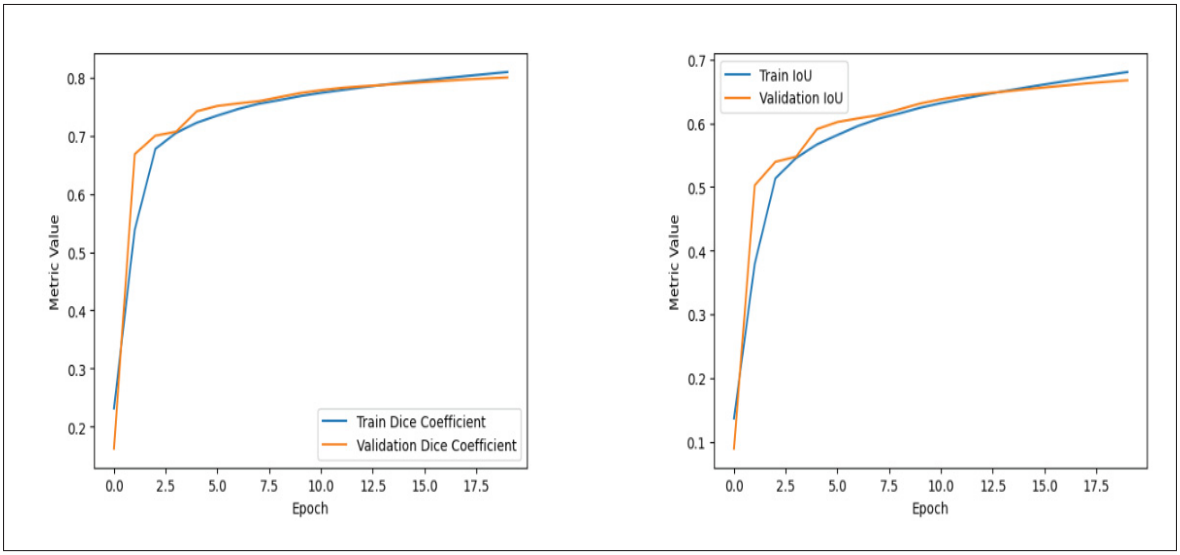

Figure 2.10 Swin-Unet results (EchoNet-Pediatric)

These metrics show a high level of agreement between model predictions and ground truth annotations, demonstrating the model's ability to accurately delimit left ventricle structure. The model training process was also effective, as reflected in the low loss value of 5%, showing that the model has learned to efficiently segment pediatric echocardiographic images while minimizing errors.

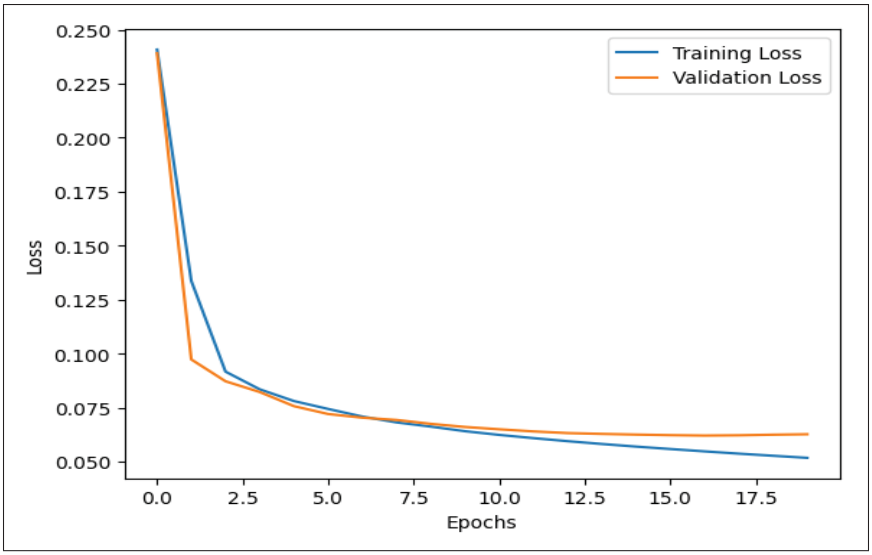

Figure 2.11 Swin-Unet results (EchoNet-Pediatric)

To best highlight the robustness of our model in both adult and pediatric databases (Reddy, Lopez, Ouyang, Zou & He (2023b)), we set up the two summary tables below. Indeed, our aim was to emphasize that the combination of convolutional neural networks with Transformers can improve the results obtained by a CNN in a segmentation task. This is illustrated in Table 1, where we compare the evaluation metrics results for the two models, U-Net and Swin U-Net, at the end of their training on the Echonet-Dynamic and EchoNet-Pediatric datasets.

Tableau 2.1 Comparison of Swin U-Net and U-Net results

Model	**Dataset**	**Dice Coefficient**	**IoU (%)**	**Binary Accuracy (%)**	**Value Error (%)**	**Hausdorff Distance**
U-Net	Echonet-Dynamic	91.66	84.63	98.45	2.09	2.15
	Echonet Pediatric	87.35	76.04	97.75	5.58	3.18
Swin U-Net	Echonet-Dynamic	88.57	78.93	97.51	4.04	2.89
	Echonet Pediatric	80.94	68.03	97.09	5.18	3.27

2.5 Conclusion

This study presents a new Swin Transformer U-Net model for the segmentation of echocardiographic images. Swin UNet model has offered several benefits over traditional models for left ventricle segmentation while integrating transformers. Our model leverages the advantages of both convolutional neural networks (CNNs) and transformers, providing a robust framework for accurate segmentation tasks, particularly in challenging cases with varied heart phenotypes. The hierarchical design enables it to process images at multiple scales, ensuring that fine details and broader structural information are captured effectively. This leads to a robust model that can generalize across various types of echocardiographic images and it could be applied in diverse clinical settings, potentially improving the diagnostic accuracy and consistency of left ventricle segmentation. Similarly, our Swin UNet model shows good segmentation accuracy, particularly due to its ability to capture long-range dependencies and contextual information in echocardiographic images. For instance, unlike other models, which rely on traditional CNN architectures, Swin UNet integrates transformerbased mechanisms, enhancing its ability to model complex spatial relationships and improving segmentation performance. This provides a greater understanding and better management of the variability of echocardiographic images.

A validation was done using the EchoNet and EchoNet-Pediatric databases, and indicated a very good performance. The automatic evaluation of echocardiography is very important for diagnosis of congenital heart disease, and it may pave the way for the automatic analysis of clinical parameters such as ejection fraction measurements. The segmentation task remains important since it outlines how ejection fraction estimations are generated. This task generates the equivalent of the manual tracing for each frame between systole and diastole, and might provide validation information about the whole echocardiographic sequence, for each beat. Moreover, since echocardiography allows to visualize the heart motion in real time, it might also provide a better understanding of cardiovascular dynamics and allow the development of more comprehensive models which incorporate the complexities of congenital heart disease (Azizmohammadi *et al.* (2022))(Azizmohammadi *et al.* (2023)). Such information could be integrated into a virtual simulator modeling complex congenital heart diseases. Future work will

involve segmentation evaluation on complex anatomies of CHD pediatric datasets. Furthermore, we will investigate our Swin-Unet segmentation model on other imaging modalities such as cardiac MRI or CT scans. This would extend the applicability of the Swin Transformer U-Net model across different imaging techniques for the study of CHD.

CHAPITRE 3

RÉSULTATS SUPPLÉMENTAIRES

Après l'entraînement du modèle Swin-Unet, nous avons procédé à l'extraction des images échocardiographiques correspondant aux deux phases clés du cycle cardiaque : la fin de diastole (ED) et la fin de systole (ES). Ces phases sont cruciales pour l'évaluation des volumes cardiaques, car elles marquent respectivement le volume maximal du ventricule gauche et le volume minimal, avant et après la contraction cardiaque. Pour calculer ces volumes, nous avons d'abord segmenté le ventricule gauche dans les images extraites de chaque phase, en utilisant le modèle Swin-Unet. Cette segmentation permet d'isoler précisément les contours du ventricule gauche, garantissant ainsi la fiabilité des mesures suivantes. Par la suite, l'aire de chaque zone segmentée a été calculée en comptant les pixels non nuls de l'image binaire segmentée. Le calcul du volume de fin de diastole (EDV) a été réalisé en intégrant les régions segmentées de l'image correspondant à la phase de diastole, où le ventricule gauche est dilaté. De même, le volume de fin de systole (ESV) a été calculé à partir des régions segmentées pendant la phase de contraction, où le ventricule est au volume le plus faible. Une fois ces volumes obtenus, nous avons utilisé ces valeurs pour prédire la fraction d'éjection (FE), un indicateur clé de la fonction cardiaque, qui permet de mesurer l'efficacité du cœur à pomper le sang.

$$FE = \frac{EDV - ESV}{EDV} \tag{3.1}$$

Comme le montre le tableau 3.1, nous avons sélectionné un échantillon de 10 images parmi les 100 testées, afin de valider notre approche. Cette évaluation met en évidence la comparaison entre les valeurs prédites de la fraction d'éjection et les valeurs réelles. Les résultats démontrent un accord important entre les mesures, soulignant la précision et la fiabilité de notre méthode dans l'estimation de la fraction d'éjection. Cette approche souligne l'importance d'une segmentation précise du ventricule gauche, car elle a un impact direct sur la précision des calculs de volume et la prédiction ultérieure de la fraction d'éjection.

Tableau 3.1 Prédictions de la fraction d'éjection par le modèle Swin-Unet

Image	**ESV**	**EDV**	**Fraction d'éjection réelle**	**Fraction d'éjection prédite**
Image 1	16.9	49.0	65.5	59.4
Image 2	16.9	48.9	65.4	59.3
Image 3	107.8	161.0	33.0	29.9
Image 4	29.0	72.9	60.2	57.9
Image 5	21.8	75.0	70.9	67.3
Image 6	41.7	92.9	55.1	49.3
Image 7	57.9	93.4	37.9	32.9
Image 8	181.1	253.0	28.4	24.8
Image 9	44.7	88.0	49.3	46.5
Image 10	319.5	353.0	9.5	8.3

En tirant parti du modèle Swin-Unet, nous avons pu obtenir une grande précision dans la segmentation de la région ventriculaire gauche, garantissant ainsi des prédictions fiables pour l'analyse fonctionnelle cardiaque.

CONCLUSION ET RECOMMANDATIONS

L'objectif de cette étude est de développer et d'évaluer une méthode de segmentation du ventricule gauche à partir d'images échocardiographiques, en utilisant le modèle Swin-Unet. Cette approche s'inscrit dans une démarche novatrice, exploitant les avantages des architectures Vision Transformer (ViT) pour améliorer la précision et l'efficacité de la segmentation dans un contexte médical spécifique. Notre étude est l'une des premières à aborder ce sujet, soulignant ainsi le potentiel d'amélioration des méthodes de segmentation et l'importance d'une telle approche pour l'analyse clinique des images échocardiographiques. Notre étude aborde les avantages des nouvelles méthodes pour améliorer les techniques de segmentation. En particulier, elle démontre comment l'intégration des architectures ViT dans la segmentation du ventricule gauche peut offrir des résultats plus précis et adaptés aux défis posés par les images échocardiographiques. Notre architecture en forme de U est particulièrement performante pour les tâches de segmentation d'images médicales. Cette structure facilite la détection d'informations contextuelles à différentes échelles, en combinant efficacement des caractéristiques de bas niveau et des caractéristiques de haut niveau. De même, les connexions résiduelles qui existent entre la partie de codage et celle de décodage permettent de conserver les informations fines lors du processus de reconstruction, ce qui est crucial pour des tâches de segmentation où la précision des contours est essentielle. L'intégration du mécanisme d'attention dans l'architecture de Swin Transformer constitue un élément fondamental de ses capacités impressionnantes en termes de segmentation d'images. Au contraire des approches classiques qui analysent chaque pixel de manière uniforme, le Swin Transformer ajuste dynamiquement son attention en se concentrant sélectivement sur différentes régions, tout en intégrant à la fois les détails locaux et les relations globales. Cette capacité à utiliser le mécanisme d'attention est particulièrement cruciale pour la segmentation du ventricule gauche afin d'identifier les zones les plus pertinentes et les plus influentes pour cette tâche. Cette approche novatrice permettra au Swin Transformer de procéder à une segmentation plus fine, même en tenant compte de la complexité des variations morphologiques, offrant

ainsi un cadre robuste et performant pour les applications de la segmentation dans le secteur de la médecine. L'intégration de ces deux architectures donnera naissance au modèle Swin Unet qui exploite à la fois les détails fins et les relations contextuelles globales. Le U-Net se distingue par sa capacité à capter les détails les plus fins grâce à ses connexions résiduelles qui conservent l'information à différentes échelles. Le Swin Transformer vient renforcer cette approche en fournissant une compréhension contextuelle globale par le mécanisme d'attention. Cette complémentarité entre ces deux composants a notamment permis de relever les défis posés par la grande variabilité des images échocardiographiques, offrant ainsi une solution robuste pour des scénarios cliniques divers.

Dans le cadre de ce projet, on a démontré la capacité du modèle Swin-Unet à effectuer la segmentation du ventricule gauche pour les images échocardiographiques. Nos résultats indiquent que Swin-Unet, grâce à son mécanisme d'attention de 'shifted window attention', offre une meilleure précision et une plus grande robustesse que les méthodes conventionnelles ainsi que plusieurs modèles d'apprentissage profond traditionnels. Pour évaluer la performance de notre modèle combiné Swin-Unet, nous avons opté pour plusieurs métriques de segmentation communes tout au long du processus d'entraînement. Le coefficient de Dice (DSC), l'Intersection over Union (IoU) ainsi que la distance de Hausdorff sont généralement les scores les plus largement utilisés pour mesurer la performance d'un modèle. Les résultats obtenus sont remarquables, car ils illustrent le potentiel des transformateurs visuels pour une meilleure analyse des images médicales. La capacité de Swin-Unet à capturer simultanément des informations locales et globales a permis une segmentation plus précise des structures anatomiques complexes présentes dans les échocardiographies. Cela pourrait entraîner des progrès cliniques substantiels, tels qu'une évaluation plus fiable de la fonction cardiaque et une meilleure planification des traitements. Bien que les résultats obtenus soient encourageants, cette étude présente certaines limitations qui méritent d'être soulignées. La nature diverse de l'ensemble de données utilisé risque de restreindre la généralisation des résultats à d'autres groupes de patients ou à des types d'images

échocardiographiques qui ne sont pas représentés dans l'ensemble de données. C'est pour cette raison que les travaux futurs porteront sur l'évaluation de la segmentation d'anatomies complexes à partir de données pédiatriques de malformations cardiaques congénitales (CHD). En outre, nous examinerons l'application de notre modèle de segmentation Swin-Unet à d'autres modalités d'imagerie telles que l'IRM cardiaque ou les tomodensitomètres. Cela permettra également d'étendre l'applicabilité du modèle Swin-Unet à différentes techniques d'imagerie pour l'étude des malformations cardiaques congénitales (CHD). Il serait aussi intéressant d'explorer de nouvelles pistes afin d'améliorer la robustesse de notre modèle lors de son utilisation. Dans ce cadre, bien que Swin Unet ait pu montrer une capacité prometteuse quant à la segmentation du ventricule gauche, on compte faire intégrer l'aspect temporaire en passant de la segmentation d'images statiques à l'analyse de séquences vidéo échocardiographiques. On vise alors à ajouter une autre variable au fil du temps dans le processus pour mieux suivre le fonctionnement de notre cœur et contrôler l'évolution du ventricule au cours du cycle cardiaque. Cette approche est cruciale pour une analyse précise et dynamique afin de superviser les changements morphologiques de notre organe. En ajoutant l'aspect temporel, il serait possible de segmenter le ventricule non seulement sur des images individuelles, mais aussi sur des séquences vidéo. C'est ainsi que les cardiologues auront la possibilité d'extraire des informations significatives à partir des visualisations et d'analyser le fonctionnement global au fil du temps. Dans ce cas, on pourra suivre la variation des indices cliniques des individus, telles que les variations de volume, la fraction d'éjection, le volume end-systolique et diastolique ainsi que la contraction et la relaxation du ventricule à chaque phase du cycle cardiaque. Cette approche offrirait une meilleure compréhension des fonctions cardiaques, ce qui améliorerait la capacité à interpréter les données échocardiographiques. Les cardiologues seront ainsi en mesure de réaliser des diagnostics plus précis et de surveiller de plus près la progression des maladies cardiaques en se basant sur ces informations visuelles et temporelles. Parallèlement à la prise en compte de la dimension temporelle, il serait également intéressant de se pencher sur les résultats de

segmentation de notre modèle. On vise à déterminer la position du grand et petit axe du ventricule gauche. Ces deux éléments constituent des indicateurs anatomiques fondamentaux souvent utilisés en cardiologie. Ces mesures servent à caractériser la performance du système cardiaque, notamment lors de l'analyse échocardiographique ou de l'IRM du cœur. Une fois que notre ventricule a été segmenté, le grand axe définit la distance maximale entre les deux points distants sur le contour. Quant au petit axe, il s'agit d'un diamètre mesuré perpendiculairement sur le grand axe. En d'autres termes, c'est le diamètre le plus petit. Une fois ces mesures prises, la taille, la forme et le volume du ventricule gauche peuvent être évalués. Ces paramètres peuvent être analysés pour étudier les performances du cœur et repérer des dysfonctionnements tels que la dilatation du ventricule.

BIBLIOGRAPHIE

Aghapanah, H., Rasti, R., Kermani, S., Tabesh, F., Banaem, H. Y., Aliakbar, H. P., Sanei, H. & Segars, W. P. (2024). CardSegNet : an adaptive hybrid CNN-vision transformer model for heart region segmentation in cardiac MRI. *Computerized Medical Imaging and Graphics*, 115, 102382.

Aubry, A. & Duong, L. (2023). Automatic evaluation of the ejection fraction on echocardiography images. *CMBES Proceedings*, 45.

Azizmohammadi, F., Navarro Castellanos, I., Miró, J., Segars, P., Samei, E. & Duong, L. (2022). Generative learning approach for radiation dose reduction in X-ray guided cardiac interventions. *Med. Phys.*, 49(6), 4071–4081.

Azizmohammadi, F., Castellanos, I. N., Miró, J., Segars, P., Samei, E. & Duong, L. (2023). Patient-specific cardio-respiratory motion prediction in X-ray angiography using LSTM networks. *Phys. Med. Biol.*, 68(2), 025010.

Beurnier, D. A. [DES de pneumologie, school = Université paris saclay,]. (2021). Circulation pulmonaire.

Bi, K., Tan, Y., Cheng, K., Chen, Q. & Wang, Y. (2022). Sequential shape similarity for active contour based left ventricle segmentation in cardiac cine MR image. *Mathematical Biosciences and Engineering*, 19(2), 1591-1608.

Bittar, D. P. (2022). Anatomie du coeur. *Le site du Collège des paramédicaux, Société Française de Cardiologie*.

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q. & Wang, M. (2021). Swin-Unet : Unet-like Pure Transformer for Medical Image Segmentation.

Cardinal, M.-H. R., Meunier, J., Soulez, G., Maurice, R. L., Thérasse, & Cloutier, G. (2005). Automatic 3D Segmentation of intravascular ultrasound images uing region and contour information. *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2005*, pp. 319–326.

Cui, H., Yuwen, C., Jiang, L., Xia, Y. & Zhang, Y. (2021). Multiscale attention guided U-Net architecture for cardiac segmentation in short-axis MRI images. *Computer Methods and Programs in Biomedicine*, 206, 106142.

Cui, Y., Jiang, C., Wu, G. & Wang, L. (2024). MixFormer : end-to-end tracking with iterative mixed attention. *IEEE Tansactions on pattern analysis and machine intelligence*, 46(6), 4129-4146.

De Santi, L. A., Meloni, A., Santarelli, M. F., Pistoia, L., Spasiano, A., Casini, T., Putti, M. C., Cuccia, L., Cademartiri, F. & Positano, V. (2023). Left ventricle detection from cardiac magnetic resonance relaxometry images using visual transformer. *Sensors*, 23(6).

Deng, K., Meng, Y., Gao, D., Bridge, J., Shen, Y., Lip, G., Zhao, Y. & Zheng, Y. (2021). TransBridge : a lightweight transformer for left ventricle segmentation in echocardiography. pp. 63–72.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. & Houlsby, N. (2021). An image is Worth 16x16 Words : transformers for image recognition at scale.

Eske, J. (2019). What is end-diastolic volume ? *Medical News Today*.

Eske, J. (2024). Ejection fraction : how is it measured ? *Hope Cristol, Stephanie Langmaid*.

Fatakdawala, H., Xu, J., Basavanhally, A., Bhanot, G., Ganesan, S., Feldman, M., Tomaszewski, J. E. & Madabhushi, A. (2010). Expectation–maximization-driven geodesic active contour with overlap resolution (EMaGACOR) : application to lymphocyte segmentation on breast cancer histopathology. *IEEE Transactions on Biomedical Engineering*, 57(7), 1676-1689.

Hoang Ngan Le, T., Luu, K., Duong, C. N., Quach, K. G., Truong, T. D., Sadler, K. & Savvides, M. (2020). Active contour model in deep learning era : a revise and review. Dans Oliva, D. & Hinojosa, S. (Éds.), *Applications of Hybrid Metaheuristic Algorithms for Image Processing* (pp. 231–260). Cham : Springer International Publishing.

Hoffman, J. I. E. & Kaplan, S. (2002). The incidence of congenital heart disease. *J. Am. Coll. Cardiol.*, 39(12), 1890–1900.

Jessica I. Gupta, M. J. S. (2022). Biologie du cœur. *Merck Canada*.

Johanne Marcotte, R. O. (2004). *Le coeur et les vaisseaux sanguins*. Bibliothèque nationale du Québec.

Koo, H. J., Lee, J.-G., Ko, J. Y., Lee, G., Kang, J.-W., Kim, Y.-H. & Yang, D. H. (2020). Automated segmentation of left ventricular myocardium on cardiac computed tomography using deep learning. *Korean Journal of Radiology*, 21(6).

Lavigne, E. (2016). Le cœur, pompe autonome mécanique et électrique. *Planète Santé*.

Liao, M., Lian, Y., Yao, Y., Chen, L., Gao, F., Xu, L., Huang, X., Feng, X. & Guo, S. (2023). Left ventricle segmentation in echocardiography with transformer. *Diagnostics*, 13(14).

Liu, X., Fan, Y., Li, S., Chen, M., Li, M., Hau, W. K., Zhang, H., Xu, L. & Lee, A. P.-W. (2021a). Deep learning-based automated left ventricular ejection fraction assessment using 2-D echocardiography. *American Journal of Physiology-Heart and Circulatory Physiology*, 321(2), H390-H399.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. & Guo, B. (2021b). Swin Transformer : Hierarchical vision transformer using Shifted windows. *CoRR*, abs/2103.14030.

Liu, Z., He, X. & Lu, Y. (2022). Combining uNet 3+ and transformer for Left ventricle segmentation via signed distance and focal loss. *Applied Sciences*, 12(18).

Magnin, E. (2023). Cœur : Anatomie, pathologies et traitements. *PasseportSanté*.

Meunier, J. (1998). Tissue motion assessment from 3D echographic speckle tracking. *Physics in Medicine Biology*, 43(5), 1241.

Michael J. Shea, T. C. (2023). Échocardiographie. *Merck Canada*.

Ouyang, D., He, B., Ghorbani, A., Yuan, N., Ebinger, J., Langlotz, C. P., Heidenreich, P. A., Harrington, R. A., Liang, D. H., Ashley, E. A. & Zou, J. Y. (2020). Video-based AI for beat-to-beat assessment of cardiac function. *Nature*, 580(7802), 252–256.

Pham, T. H. & Singh, G. D. (2024). 3D intracardiac echocardiography for structural heart interventions. *Interv. Cardiol. Clin.*, 13(1), 11–17.

Prsa, M., Saroli, T., Correa, J. A., Asgharian, M., Mackie, A. S. & Dancea, A. B. (2009). Birth prevalence of congenital heart disease. *Epidemiology*, 20(3), 466–468.

Reddy, C. D., Lopez, L., Ouyang, D., Zou, J. Y. & He, B. (2023a). Video-Based deep Learning for automated assessment of left ventricular ejection fraction in pediatric patients. *Journal of the American Society of Echocardiography*, 36(5), 482-489.

Reddy, C. D., Lopez, L., Ouyang, D., Zou, J. Y. & He, B. (2023b). Video-based deep learning for automated assessment of left ventricular ejection fraction in pediatric patients. *J. Am. Soc. Echocardiogr.*, 36(5), 482–489.

Reller, M. D., Strickland, M. J., Riehle-Colarusso, T., Mahle, W. T. & Correa, A. (2008). Prevalence of congenital heart defects in metropolitan Atlanta, 1998-2005. *J. Pediatr.*, 153(6), 807–813.

Rubeaux, M., Doris, M. K., Alessio, A. & Slomka, P. J. (2017). Enhancing cardiac PET by motion correction techniques. *Current Cardiology Reports*, 19(2).

Taboulet, P. (2024). Extrasystole ventriculaire (ESV). *E-cardiogram : de A à Z*.

Takahashi, S., Sakaguchi, Y., Kouno, N., Takasawa, K., Ishizu, K., Akagi, Y., Aoyama, R., Teraya, N., Bolatkan, A., Shinkai, N., Machino, H., Kobayashi, K., Asada, K., Komatsu, M., Kaneko, S., Sugiyama, M. & Hamamoto, R. (2024a). Comparison of vision transformers and convolutional neural networks in medical image analysis : a systematic review. *Journal of Medical Systems*, 48(1), 84.

Takahashi, S., Sakaguchi, Y., Kouno, N., Takasawa, K., Ishizu, K., Akagi, Y., Aoyama, R., Teraya, N., Bolatkan, A., Shinkai, N., Machino, H., Kobayashi, K., Asada, K., Komatsu, M., Kaneko, S., Sugiyama, M. & Hamamoto, R. (2024b). Applying masked autoencoder-based selfsupervised learning for high-capability vision transformers of electrocardiographies. *PLOS ONE*, 19(8).

Taskén, A. A., Yu, J., Berg, E. A. R., Grenne, B., Holte, E., Dalen, H., Stølen, S., Lindseth, F., Aakhus, S. & Kiss, G. (2024). Automatic detection and tracking of anatomical landmarks in transesophageal echocardiography for quantification of left ventricular function. *Ultrasound in Medicine Biology*, 50(6), 797-804.

Touil, B., Basarab, A., Delachartre, P., Bernard, O. & Friboulet, D. (2010). Analysis of motion tracking in echocardiographic image sequences : influence of system geometry and point-spread function. *Ultrasonics*, 50(3), 373-386.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M. & Luo, P. (2021). SegFormer : simple and efficient design for semantic segmentation with transformers.

Yu, C., Li, S., Ghista, D., Gao, Z., Zhang, H., Ser, J. D. & Xu, L. (2023). Multi-level multi-type self-generated knowledge fusion for cardiac ultrasound segmentation. *Information Fusion*, 92.

Zeng, Y., Tsui, P.-H., Pang, K., Bin, G., Li, J., Lv, K., Wu, X., Wu, S. & Zhou, Z. (2023). MAEF-Net : Multi-attention efficient feature fusion network for left ventricular segmentation and quantitative analysis in two-dimensional echocardiography. *Ultrasonics*, 127(106855), 106855.

Zuercher, M., Ufkes, S., Erdman, L., Slorach, C., Mertens, L. & Taylor, K. (2022). Retraining an artificial intelligence algorithm to calculate left ventricular ejection fraction in pediatrics. *J. Cardiothorac. Vasc. Anesth.*, 36(9), 3610–3616.