UQAC
Université du Québec
à Chicoutimi

ADAPTATION DU MODÈLE SEGMENT ANYTHING AUX IMAGES MÉDICALES

PAR CORIANE KEUMEDJIO TSAHAKA

MÉMOIRE PRÉSENTÉ À L'UNIVERSITÉ DU QUÉBEC À CHICOUTIMI COMME EXIGENCE PARTIELLE EN VUE DE L'OBTENTION DU GRADE DE MAÎTRISE ÈS SCIENCES EN INFORMATIQUE

QUÉBEC, CANADA

© CORIANE KEUMEDJIO TSAHAKA, 2026

RÉSUMÉ

La segmentation d'images médicales est une étape clé pour l'analyse clinique, mais elle reste difficile à automatiser de façon fiable, notamment en raison de la diversité des modalités d'imagerie, de la variabilité anatomique et du manque de données annotées. Historiquement, cette tâche a été abordée à l'aide de réseaux spécialisés entraînés sur un jeu de données donné, ce qui limite souvent leur capacité de généralisation. L'arrivée de SAM (Segment Anything Model) a marqué un tournant en proposant un modèle de segmentation guidé par des prompts, capable de segmenter des objets variés sans entraînement spécifique. Toutefois, appliqué tel quel en imagerie médicale, SAM montre des limites, ce qui a motivé plusieurs travaux récents visant à l'adapter à ce domaine.

Le projet présenté dans ce mémoire s'inscrit dans cette dynamique et propose une approche d'adaptation paramètre-efficiente de SAM, nommée H-SAM, basée sur l'intégration d'adaptateurs Hadamard dans l'encodeur d'images. Cette stratégie vise à spécialiser le modèle aux tâches de segmentation médicale tout en limitant fortement le nombre de paramètres à entraîner, afin de réduire les coûts de calcul et les besoins mémoire. Dans le cadre de ce travail, H-SAM est évalué sur plusieurs jeux de données publics représentatifs de contextes médicaux différents (lésions cutanées, nodules thyroïdiens, structures du fond d'œil). Les performances sont analysées quantitativement à l'aide des métriques IoU et Dice, et complétées par des visualisations qualitatives sous forme d'overlays et de cartes d'erreur, afin d'interpréter plus finement le comportement du modèle et la qualité des segmentations produites. Les résultats montrent que H-SAM permet d'obtenir des performances compétitives tout en n'entraînant qu'environ 0,16 million de paramètres, soit un coût d'adaptation nettement inférieur à celui des approches d'adaptation existantes.

TABLE DES MATIÈRES

Table des matières . . . . . **ii**

Liste des tableaux . . . . . **iv**

Liste des figures . . . . . **v**

Liste des abréviations . . . . . **vii**

Dédicace . . . . . **viii**

Remerciements . . . . . **ix**

I Introduction **1**

1.1 Contexte 1

1.2 Problématique 5

1.3 Méthodologie 6

1.4 Contributions 7

1.5 Structure du mémoire 7

II Etat de l'art **9**

2.1 Concepts clés 9

2.2 Modalités d'images médicales 11

2.3 Apprentissage profond 15
2.4 Apprentissage par transfert 23
2.5 Travaux spécifiques à la problématique 27
2.6 Conclusion 40

III Méthodologie 43

3.1 Architecture SAM 43
3.2 Architecture H-SAM 47
3.3 Génération des prompts 50

IV Expérimentations et résultats 52

4.1 Langage et outils de développement 52
4.2 Jeux de données 53
4.3 Paramètres d'entraînement 55
4.4 Métriques d'évaluation 56
4.5 Implémentation logicielle 58
4.6 Étude expérimentale et analyse des résultats 67
4.7 Discussions 82

V Conclusion 84

LISTE DES TABLEAUX

4.1 Répartition des données 55

4.2 Comparaison quantitative de la Phase 1 sur le dataset ISIC (Moyenne ± Écart-type). 67

4.3 Résultats sur ISIC. 74

4.4 Résultats sur TN3K. 75

4.5 Résultats sur REFUGE. 76

LISTE DES FIGURES

2.1 Exemples d'images X-ray [4, 48] 12

2.2 Exemples d'images CT [69] 12

2.3 Exemples d'images US [41] 13

2.4 Exemples d'images SPECT/CT 13

2.5 Exemples d'images IRM 14

2.6 Exemples d'images dermoscopiques 15

2.7 Exemples d'images endoscopiques 15

2.8 Architecture d'un CNN 16

2.9 Architecture de ViT 19

3.1 Architecture de SAM [43] 44

3.2 Étape 1 de H-SAM 49

3.3 Étape 2 de H-SAM 50

4.1 Quelques images et masques d'ISIC 53

4.2 Quelques images et masques du TN3k 54

4.3 Quelques images et masques du REFUGE 54

4.4 Implémentation de l'adaptateur Hadamard. 59

4.5 Implémentation de l'étape 1 de H-SAM. 60

4.6 Implémentation de l'étape 2 de H-SAM. 61
4.7 Calcul du point central (1PT). 62
4.8 Génération de points aléatoires (1PTR/3PTR). 63
4.9 Construction de la carte d'erreur. 64
4.10 Sélection du point correctif. 64
4.11 Algorithme de guidage itératif (3PTSI). 65
4.12 Génération des boîtes englobantes. 66
4.13 Pires cas 1PT. 69
4.14 Pires cas 3PTSI. 70
4.15 Pires cas BBOX_075. 71
4.16 Nombre de paramètres entraînables. 72
4.17 Visualisation en échelle logarithmique. 72
4.18 Comparaison des résultats obtenus avec la BBOX_075. 73
4.19 Analyse qualitative sur le jeu de données TN3K 78
4.20 Analyse qualitative sur la cup optique (REFUGE) 79
4.21 Analyse qualitative sur le disque optique (REFUGE) 80
4.22 Analyse qualitative sur le jeu de données ISIC 81

LISTE DES ABRÉVIATIONS

CNN Réseau de neurones convolutionnels (Convolutional Neural Network)

CT Tomodensitométrie (Computed Tomography)

GPU Graphics Processing Unit

IRM Imagerie par résonance magnétique (Magnetic Resonance Imaging)

LLM Large Language Model

MAE Masked Auto - Encoder

NLP Natural Language Processing

PEFT Parameter-Efficient Fine-Tuning

PET Tomographie par émission de positons (Positron Emission Tomography)

RGB Images en couleurs (Red - Green - Blue)

SAM Segment Anything Model

US Échographie (Ultrasound)

ViT Vision Transformer

X-ray Rayon X

DÉDICACE

À ma famille,

pour votre amour et votre soutien
tout au long de ce mémoire.

REMERCIEMENTS

Je souhaite remercier toutes les personnes qui m'ont soutenue pendant ma maîtrise et la rédaction de ce mémoire.

Je tiens à exprimer ma profonde gratitude à mon directeur de recherche, Kevin Bouchard, pour son accompagnement, son expertise et sa compréhension tout au long de ce travail.

Je remercie également Bob Menelas, ancien directeur du programme, qui m'a été d'une aide précieuse tout au long de mon parcours.

Mes remerciements vont aussi à toute l'équipe du programme pour son écoute et sa disponibilité.

Enfin, je remercie ma famille et mes amis pour leur amour et leur soutien indéfectibles.

CHAPITRE I

INTRODUCTION

1.1 CONTEXTE

La segmentation d'images est une tâche essentielle en vision par ordinateur, car elle consiste à diviser une image ou une séquence vidéo en plusieurs régions ou objets distincts [60]. Elle fournit ainsi une représentation structurée qui permet d'analyser séparément chaque partie de l'image. Grâce à cette capacité à isoler les différentes composantes d'une image, la segmentation intervient dans un large éventail d'applications, allant de la conduite autonome [9] à l'agriculture de précision [47], en passant par la réalité virtuelle [12] et l'exploration minière [57].

D'un point de vue conceptuel, cette tâche regroupe trois sous-tâches principales. La segmentation sémantique vise à attribuer à chaque pixel une classe donnée — par exemple, distinguer un piéton d'un véhicule ou d'une portion de route. La segmentation par instances s'intéresse quant à elle à différencier chaque occurrence d'un objet. Enfin, la segmentation panoptique unifie ces deux approches en fournissant simultanément une classification par pixel et une distinction entre chaque instance d'objet [47].

Dans le domaine de l'imagerie médicale, la segmentation est le plus souvent formulée comme une tâche de segmentation sémantique. Chaque pixel doit être classé comme appartenant ou non à une structure anatomique ou pathologique [53]. Elle joue ainsi un rôle essentiel dans l'analyse clinique, car elle permet de délimiter avec précision les régions d'intérêt nécessaires au diagnostic. L'identification exacte de ces structures permet d'extraire des caractéristiques pertinentes, de mener une évaluation quantitative fiable et de soutenir la prise de décision dans la détection et la prise en charge des maladies. Cette précision constitue une exigence centrale pour des applications cliniques telles que la planification chirurgicale ou la radiothérapie [66].

Avec l'essor des technologies médicales au cours des dernières décennies, les équipements d'imagerie se sont progressivement imposés comme des outils essentiels au diagnostic clinique. Aujourd'hui, des modalités telles que la tomodensitométrie (CT), l'imagerie par résonance magnétique (IRM), la tomographie par émission de positons (PET), la radiographie (X-ray) ou encore l'échographie (US) sont devenues incontournables dans la pratique hospitalière, chacune offrant une perspective particulière sur les structures internes du corps humain. À ces modalités s'ajoutent également des images RGB utilisées en biomédecine, comme les images microscopiques ou les photographies du fond d'œil, qui permettent une observation fine des tissus, des cellules ou de pathologies spécifiques [53]. Cette diversité de sources d'imagerie reflète la complexité croissante des besoins cliniques et a naturellement conduit à un intérêt accru pour le développement de méthodes de segmentation médicale.

Selon [25], les approches de segmentation des images médicales se distinguent par le niveau d'intervention humaine qu'elles requièrent. La segmentation manuelle repose entièrement sur l'expertise humaine : l'utilisateur identifie la région d'intérêt et en trace précisément

les contours. Elle fournit des annotations de référence indispensables à l'entraînement et à l'évaluation des méthodes plus avancées, mais demeure chronophage, limitée à de petits jeux de données et sujette à une forte variabilité inter- et intra-observateur. Afin de réduire cet effort tout en conservant un certain contrôle, les méthodes de segmentation semi-automatique ont été proposées comme un compromis, combinant une initialisation fournie par l'utilisateur avec des algorithmes automatiques capables d'affiner la segmentation, tout en restant sensibles aux choix initiaux et à la variabilité utilisateur. Ces limites ont progressivement conduit au développement de méthodes de segmentation entièrement automatiques, dont l'efficacité reste fortement conditionnée par la qualité des données annotées et par leur capacité à faire face aux défis propres aux images médicales, tels que la variabilité anatomique, le faible contraste, le bruit et l'hétérogénéité des protocoles d'acquisition. Dans ce cadre, les approches reposant sur une ingénierie manuelle de caractéristiques et des modèles d'apprentissage traditionnels peuvent rencontrer des difficultés de généralisation, en particulier lorsque les conditions d'acquisition ou les distributions de données diffèrent de celles observées lors de l'entraînement.

Ces difficultés ont favorisé l'adoption croissante de l'apprentissage profond, qui s'est imposé ces dernières années comme la nouvelle référence de l'état de l'art en segmentation d'images médicales [14]. Les réseaux de neurones convolutionnels (CNN) ont notamment obtenu des succès majeurs en imagerie médicale [79], grâce à leur capacité à apprendre automatiquement des représentations discriminantes directement à partir des images brutes et à produire des segmentations fines au niveau du pixel. Parmi les architectures les plus utilisées, U-Net [71] et ses nombreuses variantes occupent une place centrale, reposant sur

une architecture encodeur–décodeur enrichie de skip connections permettant de combiner des informations contextuelles de haut niveau avec des détails fins. De nombreuses extensions ont été proposées afin d'améliorer ses performances, notamment par l'introduction de mécanismes d'attention ou de modifications structurelles [59], telles que 3D U-Net [13], Res-U-Net [81], Dense-UNet [50], UNet++ [86] et nnU-Net [38]. Plus récemment, l'essor des architectures Transformer a conduit à l'émergence de modèles hybrides, comme TransUNet [8] et nnFormer [85] [75]. Malgré leurs performances élevées, ces modèles demeurent fortement spécialisés, étant généralement conçus et entraînés pour des tâches et des modalités spécifiques, ce qui limite leur capacité de généralisation à de nouveaux contextes cliniques et constitue un frein majeur à leur déploiement à grande échelle en pratique clinique [56].

C'est dans cette dynamique que s'inscrivent les modèles de fondation, des architectures conçues pour être polyvalentes et transférables entre différentes tâches. Ces modèles sont entraînés sur des ensembles de données extrêmement volumineux afin de pouvoir se généraliser à une grande variété de tâches et de domaines. Dans le domaine du traitement automatique du langage (NLP), les Large Language Models (LLMs) ont rapidement dominé le paysage, comme en témoigne le modèle fondation de Brown et al. [6], ouvrant la voie à des architectures toujours plus vastes, jusqu'au récent GPT-4 proposé par OpenAI [1]. Dans le domaine de l'image, des modèles tels que Stable Diffusion [70] et DALL·E [68] ont démontré la capacité des modèles à générer des images de haute résolution à partir de simples descriptions textuelles. En avril 2023, la publication de *SAM* [43] a constitué une avancée décisive dans le domaine de la segmentation d'images, en introduisant un modèle promptable capable de s'adapter à un large éventail d'applications sans nécessiter d'entraînement supplémentaire sur des tâches spécifiques. Ses

performances en zero-shot et few-shot, démontrées sur des images naturelles [76, 40], ont suscité un vif intérêt dans de nombreux domaines. Cependant, son application directe à l'imagerie médicale révèle rapidement ses limites. Dépourvu d'informations spécifiques au domaine biomédical, le modèle peine à s'adapter aux caractéristiques propres aux images médicales : faible contraste, frontières anatomiques ambiguës, forte variabilité inter-patients ou encore présence de lésions de petite taille. Comme le montrent plusieurs études, cette différence de distribution entre images naturelles et images médicales entraîne une baisse notable des performances du modèle, soulignant la nécessité d'adapter SAM aux spécificités du contexte médical [16, 72, 29].

1.2 PROBLÉMATIQUE

Les limites identifiées précédemment soulèvent une question centrale : comment adapter SAM pour qu'il réponde de manière optimale aux particularités des images médicales ?

Pour tenter d'y répondre, plusieurs approches fondées sur l'apprentissage par transfert ont été explorées. Certaines méthodes misent sur un réentraînement complet du modèle, tandis que d'autres privilégient des ajustements plus ciblés sur certaines parties de l'architecture. À cela s'ajoutent des approches plus légères consistant à introduire des adaptateurs dans des blocs précis pour rendre le modèle plus sensible aux caractéristiques biomédicales. Cependant, bien que ces stratégies améliorent les résultats, elles requièrent souvent des ressources importantes en termes de mémoire, de calcul et de temps.

Dans de nombreux contextes cliniques ou académiques, ces obstacles constituent un obstacle significatif à leur utilisation. Cette problématique conduit à la question précise qui guide ce mémoire : Comment adapter SAM de manière efficace, en améliorant ses performances sur les images médicales tout en réduisant significativement le coût mémoire et le temps d'entraînement nécessaires ?

1.3 MÉTHODOLOGIE

Dans ce mémoire, nous proposons d'adapter SAM à la segmentation d'images médicales en limitant les ressources nécessaires à cette adaptation. Pour atteindre cet objectif, nous utilisons une approche d'apprentissage par transfert de type Parameter-Efficient Fine-Tuning (PEFT) [32]. Cette méthode, issue des travaux en traitement automatique du langage, permet d'introduire un nombre restreint de paramètres entraînables tout en conservant l'intégralité du modèle pré-entraîné sous forme figée. Ainsi, seuls les paramètres ajoutés sont mis à jour durant l'entraînement, ce qui réduit la consommation de mémoire et le coût computationnel. Plus précisément, nous intégrons des modules d'adaptation légers, nommés adaptateurs Hadamard [10], dans l'encodeur d'images de SAM. L'objectif de cette contribution est d'évaluer dans quelle mesure l'ajout de ces adaptateurs améliore les performances de SAM sur des tâches médicales, sans recourir à un réentraînement complet du modèle ni à des modifications structurelles lourdes.

1.4 CONTRIBUTIONS

Nos expérimentations menées sur trois jeux de données médicaux issus de modalités variées démontrent l'efficacité de H-SAM, une stratégie d'adaptation paramètre-efficiente de SAM pour la segmentation d'images médicales. Il surpasse les méthodes de référence sur certains jeux de données et atteint un niveau comparable aux modèles de l'état de l'art sur d'autres, en n'entraînant qu'environ 0,16 million de paramètres, soit moins de 1% du total des paramètres du modèle original. Ce résultat fait de H-SAM une solution à la fois efficace et économe en ressources computationnelles.

1.5 STRUCTURE DU MÉMOIRE

Ce mémoire est structuré en cinq chapitres.

Le premier chapitre présente le contexte général de la segmentation d'images médicales. Il met en évidence les limites du modèle SAM dans ce domaine et introduit la problématique qui motive ce travail.

Le deuxième chapitre dresse l'état de l'art. Il établit les fondements théoriques, incluant les modalités d'imagerie et les architectures de base telles que les CNN et les ViT. Il propose également une analyse critique des méthodes de segmentation antérieures à SAM, ainsi que des différentes stratégies d'adaptation récemment publiées pour son application au domaine médical.

Le troisième chapitre expose la méthodologie retenue. Il présente d'abord l'architecture de base du modèle SAM afin d'en expliciter les principaux composants et le fonctionnement général. Il détaille ensuite l'architecture H-SAM, fondée sur l'intégration d'adaptateurs Hadamard au sein de l'encodeur d'images de SAM. Ce chapitre précise également les choix de conception relatifs à la génération des prompts.

Le quatrième chapitre regroupe l'ensemble des expérimentations réalisées. Il présente les jeux de données exploités, décrit les protocoles d'entraînement mis en œuvre, expose les résultats obtenus pour chaque jeu de données et propose une analyse critique des performances du modèle adapté.

Enfin, le cinquième chapitre conclut ce mémoire en synthétisant les principales contributions, en discutant des limites de l'approche proposée et en proposant des perspectives d'amélioration et de recherche future.

CHAPITRE II

ETAT DE L'ART

Ce chapitre est consacré aux fondements théoriques ainsi qu'aux travaux relatifs à la segmentation d'images médicales. Il introduit d'abord les concepts clés et les modalités d'imagerie couramment utilisées en milieu clinique. Il présente ensuite les architectures d'apprentissage profond employées pour la segmentation, ainsi que les principales stratégies d'apprentissage par transfert. Enfin, il se concentre sur les travaux en lien direct avec la problématique étudiée, depuis les méthodes antérieures à l'arrivée du modèle SAM jusqu'aux approches récentes visant à l'adapter au contexte médical.

2.1 CONCEPTS CLÉS

2.1.1 IMAGE MÉDICALE

Selon Pham et al. [65], une image peut être considérée comme un ensemble de mesures réalisées dans un espace bidimensionnel (2D) ou tridimensionnel (3D). Dans le contexte médical, ces mesures correspondent à des grandeurs physiques dont la nature dépend de la modalité d'acquisition : absorption des rayons X en radiographie et en tomodensitométrie,

pression acoustique en échographie, ou amplitude du signal radiofréquence en IRM. Ces mesures constituent l'information brute à partir de laquelle les structures anatomiques sont analysées.

Selon le nombre de valeurs associées à chaque position spatiale, une image médicale peut être scalaire, lorsqu'une seule mesure est enregistrée, ou vectorielle, lorsque plusieurs mesures sont associées au même emplacement. Cette distinction est importante car elle conditionne la représentation des données et les méthodes de traitement ultérieures.

Les images peuvent être acquises dans un domaine continu, par exemple sur un film radiographique, ou sous forme discrète, comme en IRM, où le signal est échantillonné et organisé en grille régulière. Dans une image discrète 2D, chaque point de mesure est appelé pixel, tandis que dans une image 3D, il est appelé voxel.

2.1.2 SEGMENTATION D'IMAGES MÉDICALES

De manière générale, la segmentation d'images médicales peut être formalisée à l'aide de la théorie des ensembles.

Soit une image I et un ensemble de contraintes de similarité C_i $(i = 1, 2, \ldots)$. L'objectif de la segmentation est de partitionner I en N régions disjointes :

$$\bigcup_{x=1}^{N} R_x = I, \qquad R_x \cap R_y = \varnothing, \quad \forall x \neq y,\ x, y \in [1, N]. \tag{2.1}$$

Chaque région R_x regroupe les pixels qui satisfont les mêmes contraintes de similarité C_i. Il en est de même pour toute autre région R_y. Les indices x et y permettent simplement de distinguer les différentes régions, et $N \geq 2$ représente le nombre de régions obtenues après segmentation [53].

Dans ce travail, nous nous sommes concentrés exclusivement sur la segmentation des images médicales 2D, car les modalités utilisées produisent des données bidimensionnelles.

2.2 MODALITÉS D'IMAGES MÉDICALES

Les modalités d'imagerie médicale désignent l'ensemble des techniques et technologies permettant de produire des représentations visuelles du corps humain [39]. Gilbert et al. [22] les regroupent en cinq grandes catégories selon leur principe physique de fonctionnement.

2.2.1 LA RADIOGRAPHIE (X-RAY)

Cette modalité utilise un rayonnement électromagnétique pour produire des images basées sur les différents niveaux d'absorption des tissus corporels. Les photons qui traversent le corps sont captés par un détecteur, puis les images présentent différentes teintes après traitement. Les tissus à haute densité, comme les os, apparaissent sous forme de zones claires, alors que les tissus à faible densité, tels que les seins et les poumons, apparaissent sous forme de zones sombres. Rapide et largement utilisée, cette technique permet de diagnostiquer des fractures, des pathologies pulmonaires ou des lésions osseuses. Les formes les plus courantes incluent la radiographie standard, la tomodensitométrie (CT), la fluoroscopie ainsi que la mammographie.

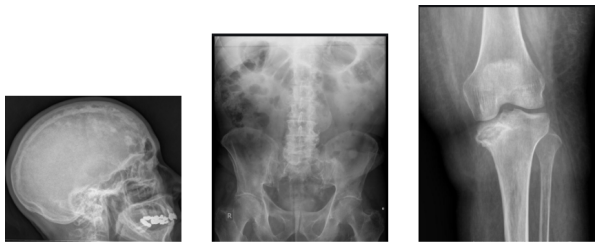


Figure 2.1 : Exemples d'images X-ray [4, 48]

Les figures 2.1 et 2.2 présentent respectivement des exemples d'images obtenues par radiographie standard et par tomodensitométrie.

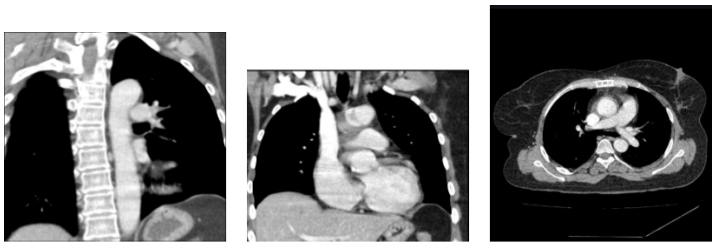


Figure 2.2 : Exemples d'images CT [69]

2.2.2 L'ÉCHOGRAPHIE (US)

Il s'agit d'une modalité qui utilise des ondes sonores de haute fréquence émises et reçues par un transducteur piézoélectrique . Les ultrasons traversent les tissus et reviennent sous forme d'échos qui varient selon la densité et la composition des structures, ce qui permet d'obtenir des images en temps réel. L'échographie est largement utilisée en obstétrique, en cardiologie et pour l'examen des organes internes car elle est non invasive et ne repose sur aucun rayonnement ionisant. Le format le plus courant est l'échographie bidimensionnelle, mais d'autres modes existent, tels que l'échographie tridimensionnelle, le Doppler pour l'étude du flux sanguin, ou encore l'échographie intracavitaire.

La figure 2.3 ci-dessous illustre plusieurs exemples d'images échographiques.



Figure 2.3 : Exemples d'images US [41]

2.2.3 L'IMAGERIE NUCLÉAIRE

Cette spécialité de l'imagerie médicale utilise des radiopharmaceutiques, qui sont des substances contenant des isotopes radioactifs, pour visualiser les fonctions physiologiques du corps. Après administration par injection, ingestion ou inhalation, le traceur s'accumule dans les organes ciblés et émet des rayonnements gamma détectés par une gamma-caméra ou un scanner spécifique. Cette technique permet d'obtenir des images bidimensionnelles planaires, des volumes tridimensionnels ou des images hybrides combinant l'anatomie et la fonction. Elle est couramment utilisée en oncologie, en cardiologie et en neurologie pour évaluer la perfusion, détecter des tumeurs ou encore surveiller l'évolution des maladies.

La figure 2.4 présente des images SPECT/CT issues de jeux de données fusionnés, obtenues dans l'étude de [2].

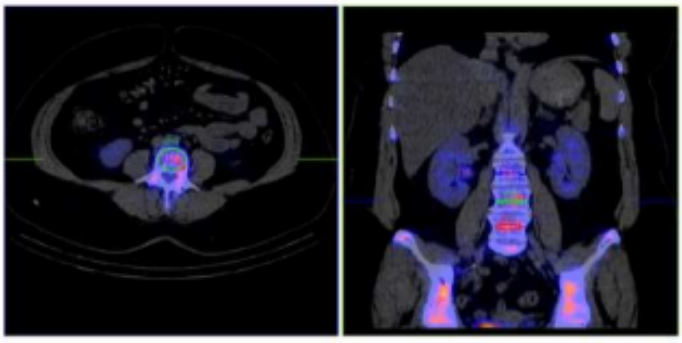


Figure 2.4 : Exemples d'images SPECT/CT

2.2.4 L'IMAGERIE PAR RÉSONANCE MAGNÉTIQUE (IRM)

L'imagerie par résonance magnétique est une modalité non invasive qui utilise un champ magnétique et des ondes radio pour produire des images détaillées des tissus internes sans recourir aux rayonnements ionisants . Elle offre un excellent contraste des tissus mous, ce qui en fait une technique de référence pour l'exploration du cerveau, de la moelle épinière et des organes abdominaux. Elle permet de produire plusieurs types d'images selon les séquences utilisées, comme les images T1 pour la vue anatomique ou les images T2 pour mettre en évidence l'inflammation ou l'œdème. D'autres séquences spécialisées, comme FLAIR, DWI, MRA ou fMRI, permettent de visualiser respectivement les lésions cérébrales, la diffusion de l'eau, les vaisseaux sanguins et l'activité cérébrale.

Des exemples d'images IRM sont présentés dans la figure 2.5.

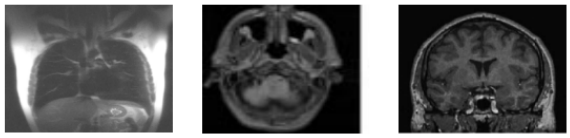


Figure 2.5 : Exemples d'images IRM

2.2.5 IMAGERIE MÉDICALE EN LUMIÈRE VISIBLE

Cette modalité repose sur l'utilisation de caméras numériques ou de capteurs optiques spécialisés qui détectent la lumière réfléchie par les tissus . Elle inclut des dispositifs tels que les endoscopes, les microscopes, les caméras ophtalmologiques et les appareils photo ou vidéo médicaux. Cette technique facilite le diagnostic clinique, la documentation des cas ainsi que la

communication entre professionnels. La dermoscopie et l'endoscopie en sont deux exemples majeurs, permettant une observation fine des tissus ou de pathologies spécifiques.

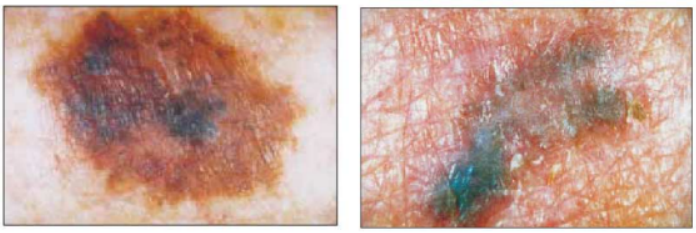


Figure 2.6 : Exemples d'images dermoscopiques

Les figures 2.6 et 2.7 illustrent respectivement quelques images obtenues par dermoscopie et par endoscopie.

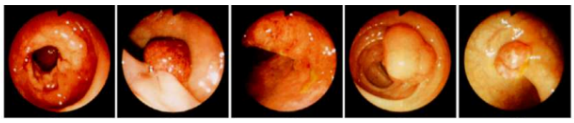


Figure 2.7 : Exemples d'images endoscopiques

2.3 APPRENTISSAGE PROFOND

L'apprentissage profond a marqué l'évolution de nombreux domaines scientifiques [3]. En segmentation médicale, les approches qui en dérivent se sont affirmées comme des outils robustes, aptes à isoler automatiquement des régions homogènes et occupant une place centrale dans les protocoles de diagnostic et d'intervention clinique [31]. Cette section introduit les architectures fondamentales sur lesquelles reposent ces approches.

2.3.1 RÉSEAU DE NEURONES CONVOLUTIFS

Les réseaux de neurones convolutionnels comptent parmi les architectures les plus représentatives de l'apprentissage profond et ont permis de nombreuses avancées en analyse d'images [53]. Inspirés du système visuel biologique et des travaux de Hubel et Wiesel sur les champs récepteurs, ils descendent du Neocognitron proposé par Fukushima en 1980 [37, 20]. Leur cadre fondamental a été établi par LeCun et al. en 1990 [45] puis perfectionné dans des travaux ultérieurs [46].

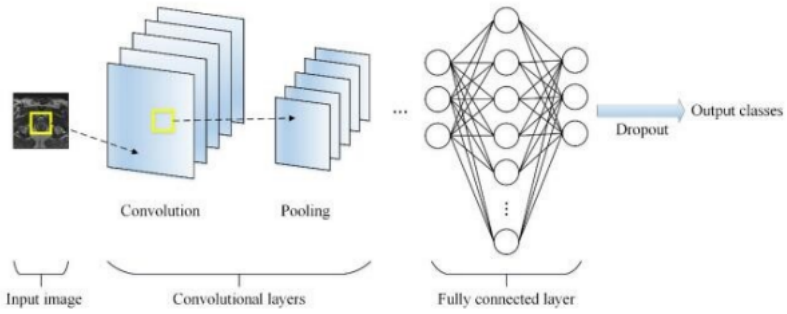


Figure 2.8 : Architecture d'un CNN

L'architecture d'un CNN, illustrée sur la figure 2.8, s'articule autour de trois grandes phases : la convolution, le pooling et les couches entièrement connectées.

La couche convolutionnelle est l'unité principale d'un CNN. Elle applique des filtres entraînables, également appelés noyaux, sur l'image d'entrée afin d'en extraire des motifs locaux tels que les bords, les textures ou des formes spécifiques. Chaque filtre parcourt systématiquement l'image et produit une carte de caractéristiques qui indique sa réponse à chaque emplacement. Le fonctionnement de cette couche est paramétré par plusieurs éléments : la profondeur, qui correspond au nombre de filtres utilisés ; le stride, qui définit le pas de

déplacement du filtre ; et le padding, qui consiste à ajouter des valeurs autour de l'entrée pour contrôler la taille de sortie. Après chaque convolution, une fonction d'activation non linéaire, généralement la ReLU, est appliquée pour augmenter la capacité de modélisation du réseau.

Le pooling est une opération d'agrégation locale qui réduit la dimension spatiale des cartes de caractéristiques tout en préservant l'information la plus pertinente. Contrairement à la convolution, le pooling ne comporte aucun poids apprenable et n'introduit donc pas de nouveaux paramètres dans le réseau. La variante la plus répandue est le max pooling, qui conserve la valeur maximale dans chaque fenêtre. Cette opération confère au réseau une meilleure robustesse aux petites translations et variations locales, car deux motifs similaires mais légèrement déplacés produiront la même activation après max pooling.

Enfin, les couches entièrement connectées intègrent les caractéristiques extraites par les couches précédentes afin de produire la prédiction finale. Dans ces couches, chaque neurone est connecté à l'ensemble des activations provenant de la couche précédente, formant ainsi un réseau dense capable de modéliser des relations complexes entre les caractéristiques apprises. La fonction d'activation appliquée est choisie en fonction de la tâche visée, par exemple ReLU pour des couches intermédiaires ou softmax pour la prédiction finale dans un problème de classification.

2.3.2 AUTO-ENCODEUR

Les autoencodeurs sont des réseaux de neurones dont l'objectif principal est de reconstruire les données d'entrée avec le minimum d'erreur possible. Pour cela, ils apprennent une

approximation de la fonction identité, en produisant une sortie $\hat{x}$ aussi proche que possible de l'entrée x.

Leur architecture est composée de deux modules complémentaires : l'encodeur et le décodeur. L'encodeur applique une transformation non linéaire $z = f_\theta(x)$ afin de projeter les données d'entrée dans une représentation latente (ou *code*), généralement de dimension inférieure. Cette étape — souvent appelée *goulot d'étranglement* — oblige le réseau à capturer les caractéristiques essentielles des données plutôt que de simplement recopier l'entrée. Le décodeur effectue ensuite la transformation inverse $\hat{x} = g_\psi(z)$ et tente de reconstruire fidèlement l'échantillon initial.

L'apprentissage consiste à optimiser simultanément les paramètres de l'encodeur et du décodeur en minimisant une fonction de coût mesurant l'erreur de reconstruction. Formulée pour un ensemble d'entraînement $(x_i)_{i=1}^n$, la fonction objectif est donnée par :

$$J_{AE}(\theta, \psi) = \sum_{i=1}^{n} L\big(x_i,\ g_\psi(f_\theta(x_i))\big),$$

où $L(\cdot)$ représente une mesure de divergence entre l'entrée et la reconstruction, telle que l'erreur quadratique moyenne ou l'entropie croisée.

Dans la tâche de segmentation médicale, les autoencodeurs ont servi d'inspiration aux architectures dédiées, comme U-Net, qui étendent leur principe d'encodeur–décodeur.

2.3.3 VISION - TRANSFORMEUR

Le Vision Transformer (ViT) a été proposé en 2021 par Dosovitskiy et al. [18] dans le prolongement des travaux fondateurs de Vaswani et al. [78], qui ont introduit le Transformer pour le traitement du langage naturel. Le succès des Transformers dans les tâches NLP a inspiré les auteurs de ViT à examiner si une approche similaire pouvait être transposée au domaine visuel.

L'idée centrale de ViT est de réduire au minimum les adaptations nécessaires pour appliquer un encodeur Transformer standard aux images. Pour cela, l'image est découpée en patches réguliers, qui sont ensuite traités comme des tokens, de la même manière que les mots dans un modèle de NLP.

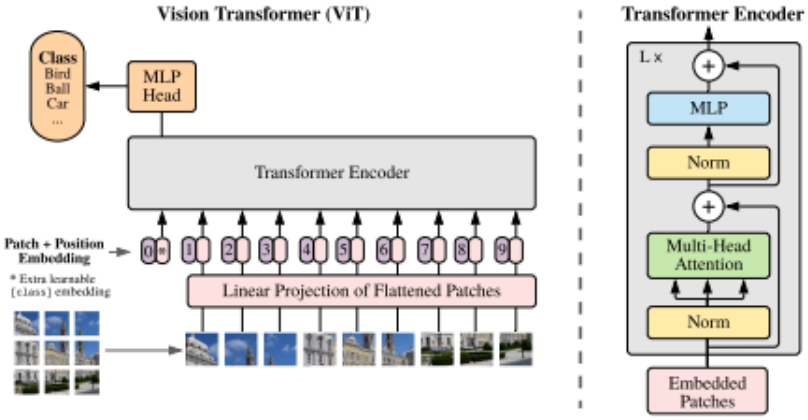


Figure 2.9 : Architecture de ViT

L'architecture ViT (Figure 2.9) s'inspire directement de l'encodeur du Transformer de Vaswani et al. Ses principaux composants sont les suivants :

1. **Découpage en patchs**

L'image d'entrée $x \in \mathbb{R}^{H \times W \times C}$ est découpée en patchs de taille fixe $P \times P$. Chaque patch est aplati en un vecteur de dimension P^2C puis projeté linéairement dans un espace latent de dimension D. Cette opération produit une séquence de $N = \frac{HW}{P^2}$ *patch embeddings*.

2. **Token spécial de classification ([CLS])**

   Un vecteur apprenable, noté [CLS], est ajouté en tête de séquence. Ce token n'est associé à aucun patch de l'image mais interagit avec eux à travers les mécanismes d'attention. Sa représentation finale servira au module de classification.

3. **Embeddings positionnels**

   La conversion de l'image en séquence fait disparaître toute information spatiale. Pour réintroduire la structure bidimensionnelle, ViT ajoute à chaque token un embedding positionnel appris. Ces vecteurs permettent au modèle de distinguer les positions relatives des patchs et facilitent la reconstruction implicite de la géométrie de l'image.

4. **Blocs Transformer**

   La séquence enrichie de ses embeddings positionnels est transmise à un empilement de L blocs Transformer. Chaque bloc contient deux sous-modules principaux :

   — **Multi-Head Self-Attention (MSA)**

   Ce module permet à chaque patch, ainsi qu'au token `[CLS]`, de tenir compte de l'information provenant de l'ensemble des autres patchs de l'image. Pour cela, chaque token de la séquence est projeté dans trois espaces distincts correspondant aux requêtes, aux clés et aux valeurs. Ces projections, notées respectivement Q, K

et V, sont obtenues au moyen de matrices linéaires apprises appliquées à la matrice des tokens Z :

$$Q = ZW^Q, \qquad K = ZW^K, \qquad V = ZW^V,$$

où W^Q, W^K et W^V sont des matrices de projection de dimension adaptée à l'espace latent. Ces vecteurs permettent de mesurer la pertinence entre les différents tokens de la séquence.

La similarité entre deux tokens est ensuite évaluée par le produit scalaire entre la requête du premier et la clé du second. Puis, cette similarité est normalisée par un facteur $\sqrt{d_k}$ et transformée par une fonction *softmax*, ce qui fournit une matrice de poids d'attention :

$$A = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right).$$

Chaque coefficient A_{ij} représente l'importance du token j pour la mise à jour du token i.

La sortie du mécanisme d'attention s'obtient alors en combinant les vecteurs de valeurs V_j selon ces poids d'importance :

$$\mathrm{Att}(Z) = A\,V.$$

Ainsi, la représentation de chaque patch est enrichie par l'information issue des régions de l'image jugées pertinentes.

Ce mécanisme est étendu au cas multi-têtes. Plutôt qu'une seule attention, le modèle en calcule plusieurs en parallèle, chacune opérant dans un sous-espace latent distinct. Pour la tête h, les projections sont :

$$Q_h = ZW_h^Q, \qquad K_h = ZW_h^K, \qquad V_h = ZW_h^V,$$

et la matrice d'attention correspondante est définie par :

$$A_h = \text{softmax}\left(\frac{Q_h K_h^\top}{\sqrt{d_k}}\right).$$

La sortie de la tête h est donc donnée par :

$$\text{Att}_h(Z) = A_h V_h.$$

Les sorties des différentes têtes sont ensuite concaténées, puis projetées dans l'espace latent initial grâce à une matrice de sortie W^O :

$$\text{MSA}(Z) = \text{Concat}\big(\text{Att}_1(Z), \ldots, \text{Att}_h(Z)\big)\, W^O.$$

Cette structure permet au modèle d'examiner simultanément plusieurs types de relations visuelles, ce qui enrichit la représentation de chaque patch et favorise l'apprentissage de motifs locaux, globaux ou structurels selon les besoins.

— **MLP feed-forward**

Après l'attention, chaque token traverse un MLP composé de deux couches linéaires séparées par une activation non linéaire (GELU). Ce module affine la représentation de chaque patch et augmente la capacité expressive du modèle.

Chaque sous-module est entouré d'une normalisation LayerNorm et d'une connexion résiduelle, ce qui stabilise l'entraînement et facilite la propagation du gradient dans les architectures profondes.

5. **Tête de classification**

   À la sortie du dernier bloc Transformer, seule la représentation du token [CLS] est conservée. Après une normalisation finale, elle est transmise à une couche linéaire qui produit la prédiction de classe. Ce mécanisme fait du token [CLS] une synthèse globale de l'image, construite progressivement par les échanges d'attention avec les différents patchs.

Dans les tâches de segmentation, la tête de classification du ViT, conçue pour des décisions globales, est écartée au profit d'un décodeur spatial adapté.

2.4 APPRENTISSAGE PAR TRANSFERT

L'apprentissage par transfert permet d'utiliser, pour une tâche cible, des connaissances acquises dans un domaine source différent, même lorsque les distributions ou les tâches ne coïncident pas. L'idée centrale consiste à exploiter des données annotées ou des représentations préalablement apprises dans un domaine connexe afin d'améliorer les performances dans un contexte où les données sont limitées [64]. Cette approche offre une alternative efficace aux

méthodes nécessitant une supervision importante et devient essentielle dans les situations où la collecte de données coûteuses ou spécialisées est difficile, comme en imagerie médicale.

Dans cette section, nous présentons tout d'abord les modèles préentraînés, puis nous décrivons les techniques d'apprentissage par transfert qui permettent de les adapter à une nouvelle tâche.

2.4.1 MODÈLES PRÉENTRAÎNÉS

Les modèles préentraînés sont des modèles dont les paramètres ont été appris à l'avance sur un grand ensemble de données avant d'être utilisés pour d'autres tâches. Cette étape leur permet d'apprendre des représentations générales qui servent ensuite de base pour différentes applications. Le préentraînement peut être supervisé, comme pour les modèles entraînés sur ImageNet [15], ou auto-supervisé, comme dans les méthodes utilisant des techniques contrastives [27] ou la prédiction de parties masquées [28]. Ces stratégies permettent d'obtenir des représentations visuelles plus robustes, qui servent souvent de point de départ pour l'entraînement de modèles en aval. Des architectures comme ResNet [26], les Vision Transformers [18], ou encore les modèles auto-supervisés tels que DINO [7] sont fréquemment utilisés comme bases préentraînées.

Les modèles de fondation (foundation models) reprennent ce principe, mais à une échelle beaucoup plus grande. Selon la définition proposée par Bommasani et al. [5], il s'agit de modèles préentraînés sur des jeux de données vastes et variés, souvent grâce à des méthodes auto-supervisées ou multimodales. Ces modèles apprennent des représentations très générales,

capables de s'adapter à un grand nombre de tâches différentes. Dans le domaine de la vision, plusieurs travaux illustrent cette tendance. CLIP [67] apprend conjointement des représentations d'images et de textes pour construire un espace commun entre les deux modalités. DINOv2 [62] propose un préentraînement auto-supervisé de grande échelle, produisant des représentations visuelles robustes et flexibles. Enfin, le Segment Anything Model, SAM [43], est un modèle de fondation orienté segmentation, préentraîné à grande échelle sur un très grand ensemble d'images et de masques afin d'apprendre une représentation adaptable à divers contextes.

2.4.2 TECHNIQUES D'APPRENTISSAGE PAR TRANSFERT

Plusieurs stratégies permettent d'adapter un modèle préentraîné à une nouvelle tâche. Une première approche consiste à utiliser le modèle comme extracteur de caractéristiques (*feature extraction*). Les représentations produites par les couches profondes sont conservées, puis un classifieur est entraîné sur la tâche cible. Cette méthode, utilisée dans les premiers travaux sur le transfert des réseaux convolutionnels [17, 73], offre un coût computationnel réduit et nécessite peu de données annotées.

Une seconde approche consiste à ajuster l'ensemble des paramètres du modèle préentraîné, ce que l'on appelle le *fine-tuning* complet. Cette stratégie permet d'adapter précisément les représentations internes du modèle et donne souvent de bons résultats lorsque les distributions source et cible sont proches [82]. Toutefois, elle devient coûteuse lorsque le modèle est de grande taille, et peut entraîner un surapprentissage lorsque les données disponibles sont limitées.

Une alternative consiste à geler une partie des couches du modèle et à n'entraîner que les couches supérieures. Cette approche a été largement utilisée après l'introduction d'AlexNet [44], car elle réduit le nombre de paramètres à optimiser tout en préservant les représentations générales apprises lors du préentraînement.

Lorsque les distributions source et cible diffèrent, des méthodes d'adaptation de domaine (*domain adaptation*) peuvent être utilisées. Ces approches cherchent à réduire l'écart entre les domaines en utilisant, par exemple, des alignements statistiques [54], des méthodes adversariales [21, 77], ou des techniques visant à minimiser explicitement la divergence entre domaines [55].

Avec l'augmentation de la taille des modèles profonds, ces techniques classiques deviennent cependant difficiles à appliquer, notamment parce que le *fine-tuning* complet nécessite une grande quantité de mémoire et peut altérer les capacités générales du modèle préentraîné. Les méthodes dites Parameter-Efficient Fine-Tuning (PEFT) ont été développées pour répondre à ces limites. Au lieu de modifier tous les paramètres du modèle, ces techniques introduisent un petit nombre de paramètres supplémentaires permettant d'adapter le modèle à une nouvelle tâche [33, 49, 52, 34].

Les adapters [33] ajoutent de petites couches entraînables insérées dans les blocs du réseau. Le prefix tuning [49] introduit des vecteurs apprenables dans les clés et requêtes du mécanisme d'attention. La méthode IA^3 [52] adapte le modèle en ajoutant des poids multiplicatifs sur certaines activations, sans modifier la structure du réseau.

Parmi les méthodes PEFT, LoRA (Low-Rank Adaptation) [34] est l'une des plus utilisées. Elle repose sur l'idée de décomposer la mise à jour d'une matrice de poids en deux matrices de faible rang. Si l'on note W_0 la matrice initiale gelée, LoRA introduit une mise à jour paramétrée par :

$$W = W_0 + BA, \tag{2.2}$$

où $A \in \mathbb{R}^{r \times d}$ et $B \in \mathbb{R}^{d \times r}$ sont des matrices de faible rang, avec $r \ll d$. Seules les matrices A et B sont entraînées, ce qui réduit considérablement le nombre de paramètres ajustés et permet d'adapter efficacement des modèles de grande taille à des jeux de données limités.

Ces méthodes PEFT sont aujourd'hui essentielles pour l'adaptation des modèles de fondation, en particulier dans les domaines où les données annotées sont peu nombreuses, comme en imagerie médicale.

2.5 TRAVAUX SPÉCIFIQUES À LA PROBLÉMATIQUE

L'arrivée de SAM [43] a marqué un tournant dans la segmentation d'images. Grâce à son préentraînement à grande échelle et à son système de prompts, SAM a été présenté comme un modèle capable de segmenter une grande variété d'objets. Dans cette section, nous présentons d'abord les méthodes d'apprentissage profond utilisées en segmentation médicale avant l'apparition de SAM. Nous décrivons ensuite les travaux qui ont appliqué SAM au domaine médical, puis ceux qui ont cherché à l'adapter spécifiquement aux contraintes de ce contexte.

2.5.1 APPROCHES ANTÉRIEURES À SAM

UNet

U-Net [71] est l'un des modèles les plus utilisés en segmentation médicale. Il repose exclusivement sur des couches convolutionnelles. Son architecture se présente sous forme d'un « U » : la première partie, appelée encodeur, réduit la taille de l'image tout en apprenant des caractéristiques importantes ; la deuxième partie, le décodeur, reconstruit une image segmentée à partir de ces caractéristiques. Les deux parties sont reliées par des connexions directes qui permettent de récupérer des détails perdus lors de la réduction de l'image.

Grâce à cette structure, U-Net parvient à segmenter des organes, des tumeurs ou des tissus, même lorsque peu d'images annotées sont disponibles. Dans de nombreuses études, il a obtenu d'excellents résultats et est devenu la base de référence dans le domaine.

U-Net a plusieurs qualités : il est simple à entraîner, s'adapte facilement à différents types d'images médicales, et produit des segmentations précises, surtout pour les petites structures. Il présente néanmoins des limites : comme il repose uniquement sur des convolutions, il peine à capturer les relations globales dans l'image, ce qui peut poser problème pour des organes de grande taille ou dont la forme varie beaucoup. Malgré cela, U-Net reste un modèle essentiel qui a influencé presque toutes les architectures de segmentation qui ont suivi.

Variantes de UNet

Plusieurs variantes de U-Net ont été proposées pour améliorer ses performances, tout en conservant la même idée générale d'un encodeur, d'un décodeur et de connexions de saut.

U-Net++ [86] modifie la manière dont les informations circulent entre l'encodeur et le décodeur. Au lieu de connexions directes, il utilise plusieurs petits blocs intermédiaires qui permettent de mieux organiser les informations à différentes échelles. Cette idée améliore la précision des contours et aide le modèle à traiter des organes ou des lésions de tailles très différentes.

Attention U-Net [61], de son côté, ajoute des modules d'attention qui permettent au modèle de se concentrer sur les régions importantes de l'image. Cette approche est particulièrement utile lorsque les structures à segmenter sont petites, peu visibles ou noyées dans un bruit important. Elle a été utilisée, par exemple, pour segmenter le pancréas ou des structures vasculaires.

D'autres variantes existent, comme ResUNet [81] ou DenseUNet [50], mais elles s'appuient toutes sur le même principe général que U-Net. Elles apportent des améliorations techniques sans modifier profondément la philosophie du modèle. Malgré leurs avantages, ces variantes restent limitées par le fonctionnement local des convolutions, ce qui explique l'intérêt croissant pour des modèles capables d'analyser l'image de manière plus globale, comme TransUNet.

TransUNet

TransUNet [8] propose une évolution importante de l'approche U-Net en intégrant un module Transformer dans l'architecture. Alors que U-Net travaille principalement sur des informations locales, les Transformers sont capables d'analyser des relations plus globales dans l'image. TransUNet combine donc ces deux idées : un encodeur basé sur un Transformer pour comprendre l'image dans son ensemble, et un décodeur convolutionnel, similaire à U-Net, pour reconstruire une segmentation détaillée.

Ce modèle a été appliqué à la segmentation d'organes abdominaux, cardiaques et cérébraux. Dans plusieurs études, il a montré de meilleures performances que les modèles purement convolutionnels, surtout lorsque la forme globale de l'organe est importante pour obtenir une bonne segmentation.

TransUNet présente cependant des limites. Les Transformers sont plus coûteux à entraîner et demandent davantage de mémoire que les CNN classiques. Ils nécessitent également, en général, plus de données annotées pour fonctionner de manière optimale. Malgré ces contraintes, TransUNet a joué un rôle majeur dans l'introduction des Transformers en imagerie médicale et a ouvert la voie à de nombreuses autres architectures hybrides.

2.5.2 SAM

Plusieurs travaux ont évalué l'utilisation directe du Segment Anything Model (SAM) [43] sur des images médicales, sans réentraînement ni modification de son architecture.

Huang et al. [36] proposent l'une des évaluations les plus vastes de SAM en segmentation médicale. Leur étude repose sur le dataset COSMOS 1050K, qui rassemble 53 jeux de données publics, 18 modalités, 84 objets anatomiques et plus d'un million d'images. Les auteurs testent six stratégies de prompting sur les deux versions de SAM, ViT-B et ViT-H : le mode Everything, un point, cinq points, cinq points positifs associés à cinq points négatifs, une boîte et une boîte associée à un point. Leurs conclusions confirment que SAM performe correctement sur les structures de grande taille et bien contrastées, notamment les organes bien délimités en radiographie ou les polypes en coloscopie, avec des DICE dépassant 90%. En revanche, ses performances chutent dès que les frontières deviennent floues, que les lésions sont petites ou que le contraste est faible. Les boîtes englobantes donnent systématiquement de meilleurs résultats que les points, et ViT-H surpasse ViT-B. L'étude montre également que les performances de SAM sont corrélées à la complexité des contours et au contraste, et que le modèle est sensible aux variations aléatoires des prompts. Enfin, les auteurs estiment que SAM peut accélérer l'annotation manuelle d'environ 25%, mais concluent qu'il reste insuffisant pour une utilisation automatique fiable en contexte clinique.

Ces résultats sont corroborés par Mazurowski et al. [58], qui réalisent une évaluation extensive de SAM sur 19 jeux de données médicaux couvrant l'IRM, le scanner, l'échographie, la radiographie et le PET. Leurs résultats montrent que les performances de SAM varient considérablement selon les tâches, allant d'un IoU de 0,11 pour la moelle épinière en IRM à 0,87 pour la hanche en radiographie. L'étude indique que les prompts par boîtes englobantes donnent de meilleurs résultats que les prompts par points, et que l'ajout de points supplémentaires en mode itératif n'améliore que marginalement les performances de SAM, le DICE passant de

56% à 64%, contrairement à d'autres méthodes interactives. Les auteurs concluent que SAM présente un potentiel pour certaines applications médicales, mais qu'il reste insuffisant pour une utilisation automatique fiable sans adaptation.

Cette variabilité des performances est également mise en évidence par He et al. [30], qui évaluent SAM sur 12 jeux de données médicaux regroupant 7 451 sujets, couvrant 10 organes et 6 modalités : IRM, scanner, radiographie, échographie, coloscopie et dermoscopie. Leurs résultats montrent que SAM est significativement moins performant que les algorithmes spécialisés comme U-Net, U-Net++ et Attention U-Net, avec des écarts de Dice allant de 0,1 à 0,5, voire de 0,6 à 0,7 dans certaines tâches. L'étude identifie plusieurs facteurs influençant les performances de SAM : la difficulté de segmentation, la dimension de l'image avec de meilleures performances en 2D qu'en 3D, la taille de la région cible où les grandes structures sont mieux segmentées, le contraste avant-plan sur arrière-plan, et la modalité d'imagerie. SAM obtient ainsi de meilleurs résultats sur radiographie, coloscopie, dermoscopie et échographie que sur IRM et scanner. Les auteurs concluent que SAM, utilisé directement sans réentraînement, n'atteint pas la précision des modèles spécialisés en segmentation médicale.

Au-delà des évaluations généralistes, d'autres travaux se concentrent sur des modalités spécifiques. C'est le cas de Roy et al. [72], qui étudient la segmentation d'organes abdominaux en scanner à partir du dataset AMOS22. Leurs expériences montrent que SAM avec des prompts par points obtient des performances nettement inférieures à l'état de l'art, avec un Dice inférieur de 20,3% à 40,9%. En revanche, les prompts par boîtes englobantes, même avec un flou modéré de 0,1, atteignent des résultats très compétitifs, proches de ceux de nnU-Net : un Dice de 77,34% avec une boîte contre 79,25% pour nnU-Net. Cette étude souligne l'importance

du choix du prompt et montre que SAM peut être efficace sur certaines tâches à condition d'utiliser une stratégie de prompting adaptée.

Enfin, Shi et al. [74] analysent SAM sur neuf benchmarks médicaux couvrant l'OCT, l'IRM, le scanner, la dermatologie, l'ophtalmologie et la radiologie. Leurs résultats montrent que SAM présente des performances de segmentation zero-shot variables selon les domaines médicaux, avec un échec complet pour certaines structures comme les vaisseaux sanguins. Sur la segmentation de lésions cutanées avec le dataset ISIC, SAM obtient un Dice de $0,73$ contre $0,92$ pour l'état de l'art. Sur les fonds d'œil, les performances varient considérablement selon les domaines, avec des Dice allant de $0,33$ à $0,58$. Les auteurs soulignent également une forte sensibilité au bruit et aux artefacts, ainsi qu'une dépendance marquée aux prompts fournis. Ils concluent que SAM ne peut pas rivaliser avec des modèles spécialisés lorsqu'il est utilisé tel quel, et qu'il doit être adapté pour atteindre des performances satisfaisantes.

Dans l'ensemble, ces travaux montrent que SAM possède des capacités intéressantes sur certaines structures bien contrastées, avec des DICE dépassant 90% en radiographie et coloscopie, mais reste très limité sur les petites structures, les contours flous et les textures homogènes, avec des DICE inférieurs à 30% sur certaines tâches. Ces limites ont motivé le développement de nombreuses méthodes visant à adapter SAM au domaine médical.

2.5.3 ADAPTATION DE SAM AUX IMAGES MÉDICALES

Afin de pallier les résultats insatisfaisants obtenus par SAM sur la segmentation d'images médicales, une approche naturelle consiste à procéder à un réentraînement ciblé sur des

données médicales. Deux grandes catégories d'adaptation sont généralement envisagées : le fine-tuning complet, qui met à jour tous les paramètres du modèle, et les méthodes de parameter-efficient fine-tuning, qui permettent d'ajuster SAM de manière plus légère et plus économique en modifiant seulement une fraction de ses paramètres.

Fine-tuning complet

Une première approche pour améliorer les performances de SAM en imagerie médicale consiste à réaliser un fine-tuning complet du modèle. Cette stratégie repose sur la mise à jour de l'ensemble des paramètres afin de réduire l'écart entre les images naturelles utilisées lors du préentraînement et les caractéristiques propres aux images médicales.

Plusieurs travaux ont montré que cette méthode permet d'obtenir des gains substantiels lorsque suffisamment de données annotées sont disponibles.

Hu et al. [35] proposent SkinSAM, une version réentraînée de SAM spécifiquement dédiée à la segmentation des lésions cutanées. À l'aide d'un réentraînement complet sur des images dermoscopiques annotées, les auteurs montrent que les performances augmentent nettement par rapport au modèle zero-shot. Le Dice score passe ainsi de 81,25% pour SAM non adapté à 88,79% après fine-tuning, soit une amélioration de 9,3%. Ce résultat illustre qu'un ajustement complet des paramètres permet au modèle de mieux capter les variations de texture et de couleur caractéristiques des lésions dermatologiques. Toutefois, cette approche reste coûteuse et nécessite un volume important d'annotations de qualité.

De manière similaire, Li et al. présentent Polyp-SAM [51] , une adaptation intégrale de SAM pour la segmentation des polypes colorectaux en endoscopie. Le modèle est réentraîné sur plusieurs bases de données publiques, ce qui lui permet d'atteindre des Dice scores supérieurs à 88% sur cinq benchmarks, avec des pointes à 92,4% sur CVC-300 et 90,6% sur Kvasir. L'amélioration est particulièrement marquée pour les polypes de petite taille ou faiblement contrastés, qui sont traditionnellement difficiles à détecter. Les auteurs comparent également deux stratégies : fine-tuner uniquement le décodeur ou fine-tuner l'ensemble des composants. Si le fine-tuning complet donne de meilleurs résultats, le gain reste limité (0 à 3%), ce qui suggère qu'une adaptation plus légère peut suffire dans certains cas.

Dans une démarche plus large, MedSAM, proposé par Ma et al. [56], réentraîne SAM sur plus d'un million de paires image-masque provenant de 11 modalités médicales différentes, incluant le CT, l'IRM, la radiographie, l'échographie et l'endoscopie. L'évaluation est menée sur 86 tâches de validation interne et 60 tâches de validation externe. Grâce à cette diversité, MedSAM atteint des performances notables : 94,4% sur les gliomes IRM T1, 94,0% sur l'hémorragie intracrânienne en CT et 98,4% pour les polypes endoscopiques, surpassant même des modèles spécialisés comme U-Net [71]. Les auteurs démontrent également que MedSAM réduit le temps d'annotation manuelle de plus de 80%. Cependant, le fine-tuning complet nécessite une puissance de calcul considérable et un volume très large d'annotations. Ce type d'approche est donc difficile à reproduire dans la plupart des contextes cliniques ou académiques.

Ces trois études montrent que le fine-tuning complet de SAM, qu'il soit appliqué à une tâche spécifique ou à une large palette de modalités, améliore significativement les performances

par rapport au modèle zero-shot. Toutefois, la puissance de calcul et le volume d'annotations nécessaires restent des obstacles majeurs, ce qui a motivé le développement de méthodes d'adaptation plus légères.

Parameter-Efficient Fine-Tuning

En réponse au coût élevé du fine-tuning complet, de nombreux travaux ont exploré des stratégies reposant sur le PEFT.

Parmi ces approches, Med-SA [80] repose sur l'insertion d'adaptateurs dans l'encodeur d'images. Ces petits modules entraînables sont ajoutés tout en laissant le reste du modèle figé. L'originalité de cette méthode réside dans deux mécanismes complémentaires. Le premier, SD-Trans, vise à adapter SAM, conçu pour des images 2D, aux données volumétriques comme le CT et l'IRM. Plutôt que de traiter chaque coupe indépendamment, ce module introduit une branche dédiée à l'extraction des corrélations entre coupes successives, capturant ainsi la troisième dimension. Le second mécanisme, HyP-Adpt, ajuste dynamiquement les représentations internes du modèle en fonction des prompts fournis par l'utilisateur. Concrètement, les embeddings des prompts sont utilisés pour générer des poids qui modulent les caractéristiques extraites par les adaptateurs. Testé sur 17 tâches couvrant 5 modalités, Med-SA obtient des résultats supérieurs à ceux de SAM en zero-shot et rivalise avec certaines méthodes de fine-tuning complet. Cependant, au-delà de l'ajout d'adaptateurs, la méthode introduit deux mécanismes supplémentaires qui alourdissent son implémentation, la faisant sortir du cadre du PEFT pur.

SAMed [83] utilise LoRA pour adapter SAM à la segmentation médicale. Des modules LoRA sont insérés dans l'encodeur d'images, ce qui permet de ne mettre à jour que 5,25% des paramètres du modèle (18,81 millions contre 358 millions pour SAM complet). Contrairement à Med-SA qui cible uniquement l'encodeur, SAMed ajuste également l'encodeur de prompts et le décodeur de masques. Cette stratégie plus globale améliore la sensibilité du modèle à la structure anatomique. Sur le dataset Synapse (scanners abdominaux), SAMed atteint un Dice score de 81,88%, surpassant ainsi des modèles spécialisés comme TransUNet (77,48%), SwinUNet (79,13%) et nnU-Net (86,33%). Toutefois, le réentraînement de l'encodeur de prompts et du décodeur sort du cadre du PEFT pur, car ces composants sont entièrement mis à jour et non simplement adaptés via LoRA.

Zhong et al. [84] proposent Conv-LoRA, une méthode qui étend LoRA en intégrant des opérations de convolution légères au sein de son goulot d'étranglement. Cette approche part du constat que l'encodeur ViT de SAM souffre de deux limitations inhérentes : d'une part, l'absence de biais inductifs liés à l'image, caractéristiques des architectures convolutionnelles, et d'autre part, un préentraînement centré sur la prédiction de masques binaires qui nuit à sa capacité à extraire des informations sémantiques de haut niveau. Conv-LoRA adresse ces deux limitations simultanément : LoRA rétablit la capacité du modèle à apprendre des représentations sémantiques, tandis que les convolutions injectent des biais inductifs locaux directement dans l'encodeur ViT. Pour déterminer l'échelle optimale à laquelle injecter ces biais, la méthode s'appuie sur un mécanisme de Mixture-of-Experts (MoE) qui sélectionne dynamiquement, pour chaque entrée, l'expert spécialisé dans l'échelle de caractéristiques la plus pertinente. Évaluée sur des benchmarks de segmentation médicale, Conv-LoRA surpasse

les autres méthodes PEFT comparées sur la quasi-totalité des tâches, avec notamment un Dice de 85,7% sur ISIC 2017 contre 84,9% pour LoRA, et un score S de 92,0% sur Kvasir contre 91,2% pour LoRA, en ne modifiant que 0,63% des paramètres du modèle. Les auteurs montrent également que l'introduction de biais inductifs améliore l'efficacité du modèle en régime de faibles données. Toutefois, les opérations d'interpolation liées au MoE engendrent un surcoût en temps d'entraînement non négligeable, Conv-LoRA étant environ 1,5 fois plus lent que LoRA.

Dans la même lignée, Khazem [42] propose TopoLoRA-SAM, qui combine LoRA et un adaptateur convolutionnel léger pour la segmentation binaire de structures fines. L'encodeur ViT-B de SAM est entièrement figé, des modules LoRA étant injectés dans les couches feed-forward de chaque bloc Transformer, tandis qu'un adaptateur convolutionnel depthwise-separable opère sur les cartes de caractéristiques haute résolution afin d'affiner les détails spatiaux fins avant le décodage. Une originalité de cette approche réside dans l'intégration d'une fonction de perte topologique, la clDice, qui optimise explicitement le chevauchement des squelettes et encourage la connectivité des structures tubulaires, particulièrement importante pour la segmentation de vaisseaux rétiniens. Évaluée sur cinq benchmarks couvrant la segmentation de vaisseaux rétiniens, de polypes et d'imagerie SAR, TopoLoRA-SAM atteint le meilleur Dice moyen global (0,735) et le meilleur Dice moyen sur les données rétiniennes (0,595), tout en n'entraînant que 5,2% des paramètres du modèle. Toutefois, la perte clDice introduit une sensibilité au poids de régularisation : une valeur trop élevée peut déstabiliser l'entraînement, tandis qu'une valeur trop faible apporte un bénéfice négligeable.

Feng et al. [19] proposent une approche PEFT pour les scénarios avec très peu de données annotées. Leur méthode combine LoRA avec un module de synthèse de données guidée par quelques exemples. Validée sur BraTS (tumeurs cérébrales) et Synapse (scanners abdominaux), l'approche atteint respectivement 77,6% et 68,2% de Dice avec seulement 10 et 5 images annotées. La limite principale est l'ajout du module de synthèse de données, qui sort du cadre du PEFT pur et introduit une étape supplémentaire dans le pipeline.

Contrairement aux autres approches, SAM-Med2D [11] représente une adaptation plus ambitieuse de SAM. Les auteurs partent du constat que le modèle, préentraîné sur des images naturelles, ne parvient pas à généraliser correctement vers des modalités médicales dont les propriétés visuelles diffèrent fortement (faible contraste, textures homogènes, absence de couleur, structures anatomiques complexes). Pour réduire cet écart, ils réentraînent SAM sur un corpus médical très large comprenant plusieurs millions d'images et de masques issus de nombreuses modalités, parmi lesquelles le CT, l'IRM, la radiographie, l'échographie et l'endoscopie. Sur le plan architectural, SAM-Med2D repose sur l'ajout de modules adaptateurs au sein de l'encodeur d'images. Les auteurs réentraînent également l'encodeur de prompts ainsi que le décodeur de masques. SAM-Med2D dépasse largement les performances de SAM en mode zero-shot et obtient des scores robustes sur plusieurs benchmarks indépendants. Cependant, cette méthode combine l'ajout d'adaptateurs avec un fine-tuning à grande échelle, sortant ainsi du cadre du PEFT pur : l'encodeur de prompts et le décodeur sont entièrement réentraînés, et le volume de données nécessaire reste considérable.

Ces méthodes illustrent la diversité des stratégies d'adaptation légère de SAM au domaine médical. Si elles améliorent toutes significativement les performances par rapport au modèle

zero-shot, aucune ne constitue du PEFT pur au sens strict. Toutes intègrent des mécanismes supplémentaires — adaptation 3D, réentraînement partiel, convolutions multi-échelles, supervision topologique ou synthèse de données — qui, bien qu'efficaces, introduisent une complexité additionnelle s'éloignant de la légèreté théorique du PEFT. Ce constat motive le développement d'approches plus sobres, capables d'améliorer les performances de SAM sur les images médicales tout en restant dans un cadre d'adaptation véritablement minimal.

2.6 CONCLUSION

Au fil de ce chapitre, nous avons retracé le cheminement des méthodes de segmentation d'images médicales en partant des notions fondamentales nécessaires pour définir ce que recouvrent une image médicale et sa segmentation. L'examen des différentes modalités d'imagerie a permis de mettre en évidence la diversité des données médicales, tant du point de vue des principes physiques que des caractéristiques visuelles, ainsi que les difficultés que cette hétérogénéité introduit pour les méthodes de segmentation.

Nous avons ensuite montré comment les architectures d'apprentissage profond, qu'il s'agisse des réseaux convolutionnels, des autoencodeurs ou des modèles de type Transformer, se sont progressivement imposées comme des outils centraux pour traiter cette complexité. Ces architectures ont permis des avancées significatives en segmentation d'images médicales, mais l'analyse de la littérature souligne qu'elles sont le plus souvent conçues pour des tâches et des modalités spécifiques, ce qui limite leur réutilisation directe dans d'autres contextes cliniques.

L'apprentissage par transfert a alors été introduit comme un prolongement naturel de ces approches, en offrant la possibilité d'exploiter des modèles préentraînés afin de réduire les besoins en données annotées et en temps d'entraînement. Parmi ces modèles préentraînés, les modèles de fondation se distinguent par un entraînement à très grande échelle sur des données variées, visant une réutilisation large sur différentes tâches. C'est dans cet ordre d'idées que s'inscrit SAM, un modèle de fondation pour la segmentation, préentraîné à grande échelle et conçu pour être adapté à une large variété de tâches.

Après avoir examiné les approches antérieures, nous avons analysé les travaux consacrés à SAM, ainsi que les tentatives visant à l'adapter au contexte médical. Ces travaux mettent en évidence que, si SAM propose un cadre générique puissant pour la segmentation, son application aux images médicales ne peut être envisagée sans une adaptation spécifique, en raison des particularités des modalités d'imagerie et des contraintes propres aux données médicales. Les stratégies d'adaptation explorées, qu'elles reposent sur le fine-tuning complet ou sur des approches de type PEFT, améliorent significativement les performances de SAM sur les images médicales. Cependant, elles introduisent toutes des mécanismes supplémentaires — réentraînement partiel, modules additionnels, synthèse de données, supervision topologique — qui alourdissent leur implémentation et s'éloignent d'un cadre d'adaptation véritablement minimal.

Ce constat conduit naturellement au point central de ce mémoire : comment tirer parti des capacités de généralisation d'un modèle de segmentation générique et préentraîné tel que SAM, tout en l'adaptant efficacement au contexte des images médicales, sans recourir à des

stratégies d'entraînement lourdes ou coûteuses. La suite de ce travail s'inscrit directement dans cette perspective.

CHAPITRE III

MÉTHODOLOGIE

Nous avons dressé, dans le chapitre précédent, un panorama des méthodes de segmentation d'images médicales, depuis les approches antérieures à l'avènement de SAM jusqu'aux travaux visant à l'adapter au contexte médical.

Dans cette continuité, l'architecture de SAM est d'abord rappelée, car elle constitue la base de l'approche proposée. L'architecture H-SAM est ensuite présentée. Cette dernière s'inscrit parmi les stratégies d'adaptation de SAM au domaine médical. Les choix retenus pour la génération des prompts sont enfin décrits.

3.1 ARCHITECTURE SAM

SAM est un modèle de segmentation d'images développé par Meta AI dans le but de constituer un modèle de fondation pour la segmentation visuelle. Contrairement aux approches classiques spécialisées pour une tâche précise, SAM est conçu pour segmenter n'importe quel objet dans une image à partir de simples prompts fournis par l'utilisateur. Ces prompts peuvent être de différentes natures : des points, des boîtes englobantes, des masques ou encore du texte.

Le modèle a été entraîné à très grande échelle sur le jeu de données SA-1B, composé de plus de 11 millions d'images haute résolution issues de scènes naturelles variées, et de plus d'un milliard de masques annotés automatiquement. Les images proviennent exclusivement de sources sous licence et respectant la vie privée, avec un floutage systématique des visages et des plaques d'immatriculation. Cette diversité des données confère à SAM une capacité de généralisation notable, lui permettant d'être appliqué à de nombreux domaines sans réentraînement préalable.

SAM est décliné en trois versions principales : SAM-B (Base), SAM-L (Large) et SAM-H (Huge). Ces variantes se distinguent essentiellement par la taille de leur encodeur d'images, le nombre de paramètres et le coût computationnel, tout en partageant une architecture fonctionnelle identique.

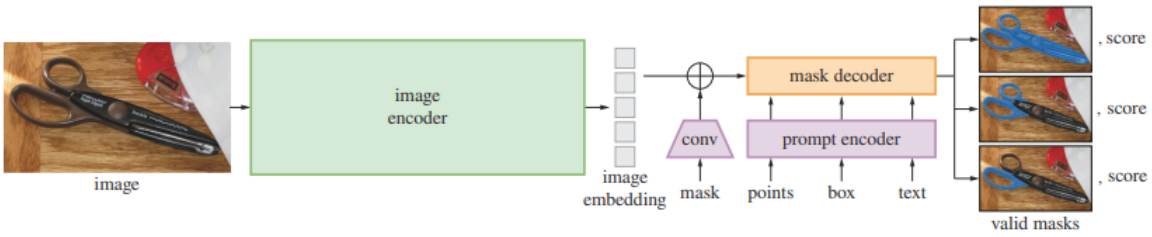


Figure 3.1 : Architecture de SAM [43]

Globalement, l'architecture de SAM, telle qu'illustrée à la figure 3.1, repose sur trois modules principaux qui interviennent de manière séquentielle : l'encodeur d'images, l'encodeur de prompts et le décodeur de masques. Le fonctionnement global du modèle consiste à extraire une représentation riche de l'image, puis à l'associer aux informations issues du prompt afin de générer un ou plusieurs masques de segmentation pertinents.

— **L'encodeur d'images**

L'encodeur d'images constitue la composante la plus lourde du modèle SAM. Reposant sur une architecture Vision Transformer (ViT), il est pré-entraîné par apprentissage auto-supervisé selon la méthode du Masked Autoencoder (MAE). Son rôle est de projeter une image haute résolution de 1024×1024 pixels vers une représentation dense de bas niveau (embedding) de $64 \times 64 \times 256$. Cet encodage capture les caractéristiques sémantiques et spatiales essentielles à la segmentation.

Cette étape n'est effectuée qu'une seule fois par image, permettant ensuite une interaction en temps réel avec le décodeur. Toutefois, l'encodeur concentrant la grande majorité des paramètres, son fine-tuning complet s'avère particulièrement coûteux en ressources GPU. Dans notre approche, cet encodeur est donc complètement figé. Cette décision impose l'introduction de mécanismes d'adaptation légers pour pallier le fossé sémantique entre les images naturelles du jeu de données SA-1B et la complexité des structures anatomiques présentes dans nos différents jeux de données médicaux.

— **L'encodeur de prompts**

L'encodeur de prompts a pour fonction de transformer les instructions fournies par l'utilisateur en représentations numériques exploitables par le modèle. Cette étape s'avère déterminante car elle guide la segmentation vers l'objet d'intérêt spécifique en traitant les informations selon leur nature. Pour les prompts épars, tels que les points et les boîtes englobantes, le modèle combine leurs coordonnées spatiales avec des vecteurs de caractéristiques appris, ou embeddings, afin de les transformer en tokens qui sont ensuite injectés directement dans les mécanismes d'attention du décodeur. En ce qui concerne les prompts denses, comme les masques, ceux-ci sont projetés via des

opérations de convolution pour atteindre une résolution de 64×64, identique à celle de l'embedding de l'image. Ils sont ensuite fusionnés par sommation avec ce dernier, enrichissant ainsi l'information visuelle par des contraintes spatiales préalables.

Le rôle de cet encodeur est d'assurer la cohérence entre l'intention de l'utilisateur et le contenu visuel brut extrait par l'encodeur d'images. Dans le cadre de la segmentation médicale, où les contrastes entre tissus sains et pathologiques peuvent être faibles, cette étape conditionne fortement la capacité du modèle à isoler précisément la zone d'intérêt malgré le bruit ou les artefacts présents dans les images.

— **Le décodeur de masques**

Le décodeur de masques a pour mission de générer la segmentation finale en fusionnant l'embedding d'image et les tokens issus de l'encodeur de prompts. Cette structure repose sur un Two-way Transformer block léger qui orchestre des mécanismes d'attention croisée bidirectionnels entre les caractéristiques de l'image et les informations de guidage. Pour transformer ces représentations latentes en un masque à haute résolution, le décodeur utilise des couches de convolutions transposées permettant de projeter l'information de l'espace d'embedding (64×64) vers la taille de sortie souhaitée.

Une particularité essentielle de ce module réside dans sa capacité à produire simultanément plusieurs propositions de masques pour un même prompt. Ce mécanisme est crucial pour résoudre les ambiguïtés structurelles, par exemple lorsqu'un point unique peut désigner à la fois une sous-partie d'une lésion ou la lésion dans son intégralité. Chaque prédiction est accompagnée d'un score de confiance estimé, fondé

sur l'indice d'Intersection over Union (IoU), permettant d'ordonner la pertinence des résultats. Cette architecture garantit une segmentation robuste, particulièrement indispensable en imagerie médicale où les frontières entre tissus sains et pathologiques sont souvent diffuses, rendant l'interprétation d'un prompt imprécis plus complexe.

Grâce à cette architecture modulaire et à son entraînement à très grande échelle, SAM démontre des capacités zero-shot étendues, facilitant son déploiement dans des domaines variés, dont le milieu médical. Néanmoins, le fossé sémantique séparant les scènes naturelles des structures cliniques nécessite l'emploi de stratégies d'adaptation légères pour garantir sa précision. C'est dans cette perspective que s'inscrit notre architecture H-SAM, dont nous détaillons la structure et la stratégie d'adaptation dans la section suivante.

3.2 ARCHITECTURE H-SAM

L'architecture H-SAM repose sur une stratégie d'adaptation paramètre-efficiente de SAM, qui vise avant tout à rendre possible son utilisation sur des tâches de segmentation spécifiques sans avoir à recourir à un fine-tuning complet, souvent coûteux en calcul et en mémoire. Dans cette approche, l'architecture fonctionnelle globale de SAM est conservée, et l'adaptation est volontairement limitée à certaines composantes bien ciblées du réseau, de manière à préserver les capacités de généralisation acquises lors du préentraînement tout en réduisant les risques de surapprentissage. Parmi les méthodes PEFT disponibles, le Hadamard Adapter a été retenu sur la base des résultats obtenus par Chen et al. [10], qui démontrent que cette méthode atteint 99,4% des performances du fine-tuning complet en moyenne sur le benchmark GLUE, tout en ne mobilisant que 0,033% des paramètres du modèle. La comparaison avec les autres méthodes

PEFT est particulièrement éloquente : sur RoBERTa-large, le Hadamard Adapter égale LoRA avec un score moyen de 88,6%, en utilisant un nombre de paramètres sept fois inférieur (0,03% contre 0,23%). Sur BART-large, il surpasse nettement les méthodes comparées avec un score moyen de 86,6%, en mobilisant 425 fois moins de paramètres que BARTen-FbT (0,02% contre 8,52%). Ces résultats confirment que le Hadamard Adapter constitue la méthode PEFT la plus efficiente en termes de paramètres tout en maintenant des performances compétitives, ce qui en fait un choix particulièrement adapté aux contraintes de calcul et de mémoire propres à la segmentation médicale.

La stratégie Hadamard mise en œuvre dans H-SAM s'organise autour de deux étapes d'entraînement successives, pensées comme un processus progressif plutôt que comme une adaptation brutale du modèle. La première étape (Figure 3.2) consiste à figer l'ensemble des paramètres du modèle SAM pré-entraîné et à n'entraîner que le décodeur de masques sur la tâche cible, ce qui permet d'adapter les sorties du modèle à la distribution de la tâche considérée avec un coût d'apprentissage très limité, tout en conservant intactes les représentations internes produites par l'encodeur d'images.

La seconde étape (Figure 3.3), qui constitue le cœur de la stratégie Hadamard, introduit des adaptateurs Hadamard directement à la sortie des mécanismes d'auto-attention dans les différentes couches de l'encodeur d'images, l'idée étant d'intervenir là où le modèle encode les dépendances globales les plus structurantes.

Concrètement, chaque adaptateur reçoit en entrée la sortie $A \in \mathbb{R}^d$ d'un bloc d'auto-attention et lui applique une transformation élément par élément : chaque dimension de A est

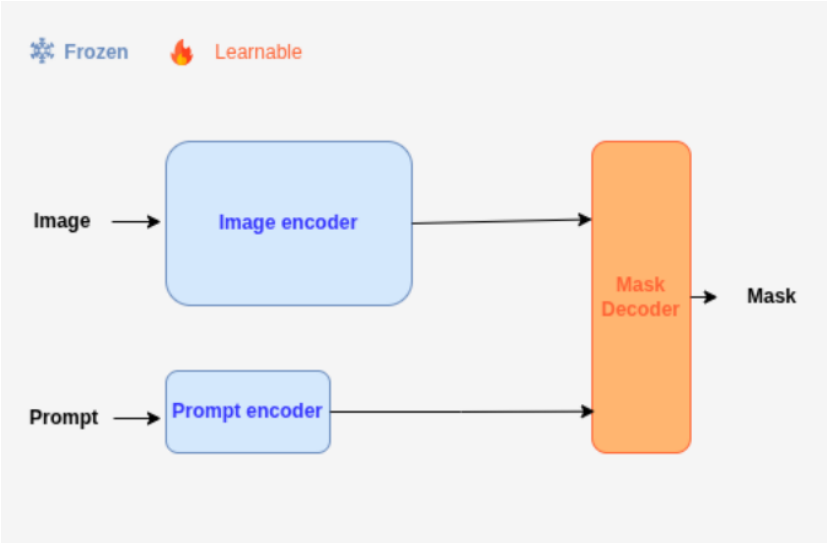


Figure 3.2 : Étape 1 de H-SAM

multipliée par un poids appris, puis décalée par un biais appris. Cette opération, connue sous le nom de produit de Hadamard, produit une nouvelle représentation A' de même dimension que l'entrée :

$$A' = W \odot A + b$$

où $W \in \mathbb{R}^d$ est le vecteur de pondération appris, $b \in \mathbb{R}^d$ le vecteur de biais appris et $\odot$ désigne le produit de Hadamard. Les paramètres sont initialisés avec $W = 1$ et $b = 0$, de sorte qu'au début de l'entraînement, l'adaptateur n'altère pas le comportement initial du modèle et que l'adaptation parte d'un état neutre.

Afin d'assurer une adaptation stable lors de cette seconde étape, la couche de normalisation placée immédiatement après l'attention multi-tête au sein de l'encodeur d'images est également dégelée, car l'introduction des adaptateurs modifie la distribution des activations internes, et laisser les paramètres de normalisation figés pourrait freiner la convergence ou dégrader les performances. En revanche, l'ensemble des autres paramètres du modèle demeure entièrement

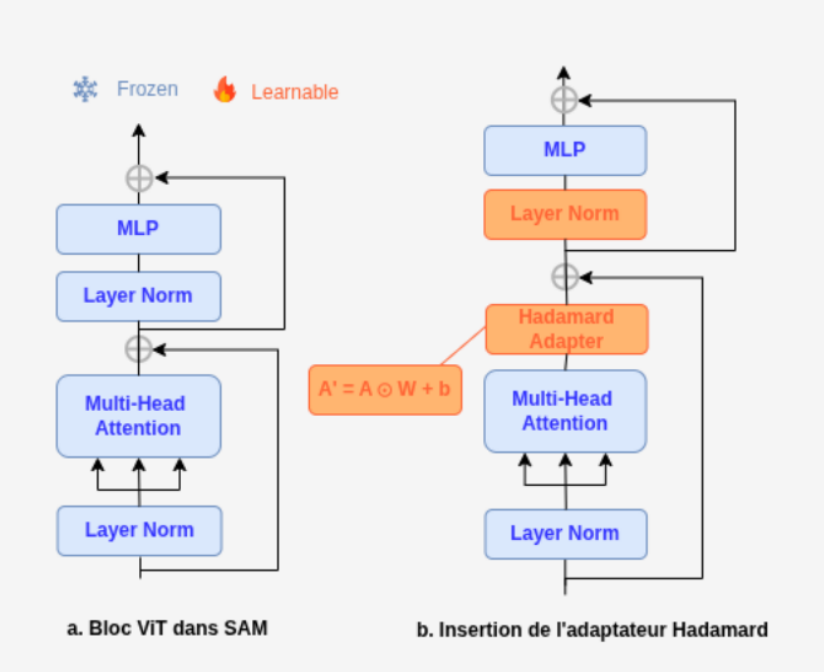


Figure 3.3 : Étape 2 de H-SAM

figé, ce qui garantit que l'adaptation reste strictement paramètre-efficiente tout en ciblant les composantes les plus influentes du réseau.

Ainsi, H-SAM se caractérise par une stratégie d'adaptation en deux étapes qui combine une première spécialisation des sorties et une modulation ciblée des représentations internes via les adaptateurs Hadamard, permettant de conserver les bénéfices du pré-entraînement à grande échelle de SAM tout en proposant un cadre d'adaptation plus léger, plus stable et mieux adapté aux contraintes propres à la segmentation médicale.

3.3 GÉNÉRATION DES PROMPTS

La performance des modèles de segmentation interactive est corrélée à la qualité des signaux d'entrée. Pour ce travail, nous nous sommes inspirés des travaux proposés dans MSA

(Medical SAM Adapter) [80], tout en développant notre propre protocole de test visant à déterminer si le modèle nécessite une précision chirurgicale pour être efficient ou s'il tolère une approximation significative.

Dans cette optique, la première étape de notre évaluation porte sur les prompts de type point, déclinés en quatre configurations. Nous cherchons d'abord à isoler l'impact de la précision à travers le cas le plus simple : le point unique. Nous comparons pour cela un point placé au centroïde du masque à un point placé aléatoirement, simulant la variabilité d'un clic manuel. Nous étendons ensuite cette analyse aux configurations à trois points, pour voir si davantage de points peut compenser un placement moins précis. Nous comparons un mode statique, où les points sont placés de manière approximative, à un mode interactif, où chaque point est ajusté de façon réactive pour corriger les erreurs du modèle.

Enfin, nous évaluons les boîtes englobantes selon la même logique. Nous comparons une boîte bien ajustée (BBOX 0,75) à une boîte plus approximative (BBOX 0,5), pour vérifier si le modèle peut identifier correctement la structure cible même sans un cadrage parfait.

CHAPITRE IV

EXPÉRIMENTATIONS ET RÉSULTATS

Le chapitre précédent était consacré à la méthodologie adoptée dans ce travail, avec la description de l'architecture de SAM, de l'approche proposée H-SAM, ainsi que des stratégies de génération des prompts.

Dans ce chapitre, nous détaillons le protocole expérimental mis en œuvre ainsi que l'analyse des résultats obtenus. Nous présentons d'abord l'environnement de développement et les jeux de données, puis les paramètres retenus pour l'entraînement et l'évaluation du modèle. Nous décrivons également l'implémentation logicielle des deux étapes de H-SAM et des fonctions de génération de prompts. Enfin, les performances de H-SAM sont passées au crible via une analyse quantitative et des visualisations qualitatives.

4.1 LANGAGE ET OUTILS DE DÉVELOPPEMENT

L'ensemble des expérimentations a été réalisé en Python 3.12 au sein de l'environnement de développement PyCharm. Les architectures des modèles ainsi que les protocoles d'entraînement et d'évaluation reposent sur la bibliothèque PyTorch. La variante SAM-Huge a été retenue

afin d'assurer une comparaison cohérente avec les travaux existants dans la littérature, qui s'appuient majoritairement sur cette version. La charge de calcul qu'elle implique dépassant les capacités des ressources locales disponibles, les calculs ont été déportés sur le cluster Fir (anciennement Cedar) de l'Alliance de recherche numérique du Canada, qui a fourni l'infrastructure GPU nécessaire à l'entraînement et à l'évaluation du modèle.

4.2 JEUX DE DONNÉES

Nous avons sélectionné trois jeux de données publics afin d'évaluer l'approche H-SAM dans des contextes médicaux variés, en nous alignant sur les choix effectués dans Med-SA [80] qui couvre les mêmes modalités.

Le premier jeu de données, ISIC 2016 [24], est composé d'images dermatoscopiques utilisées pour la segmentation de lésions cutanées. La Figure 4.1 présente quelques exemples d'images ainsi que leurs masques correspondants.

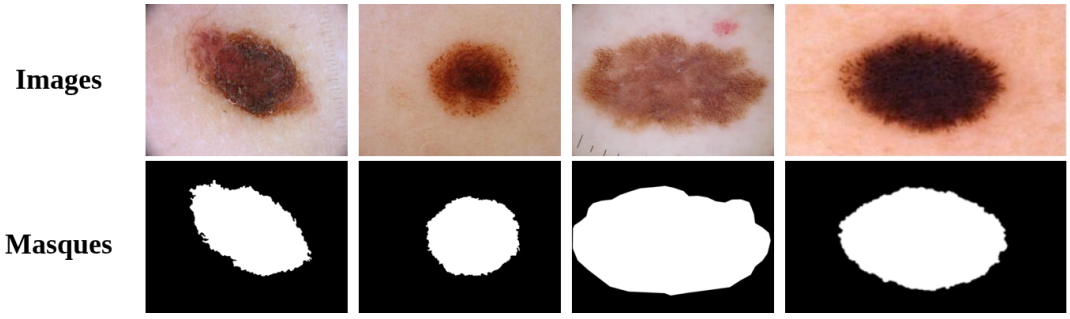


Figure 4.1 : Quelques images et masques d'ISIC

Le deuxième jeu de données, TN3k [23], est dédié à la segmentation des nodules thyroïdiens à partir d'images échographiques. La Figure 4.2 permet d'illustrer la variabilité des images ainsi que la localisation des nodules à segmenter.

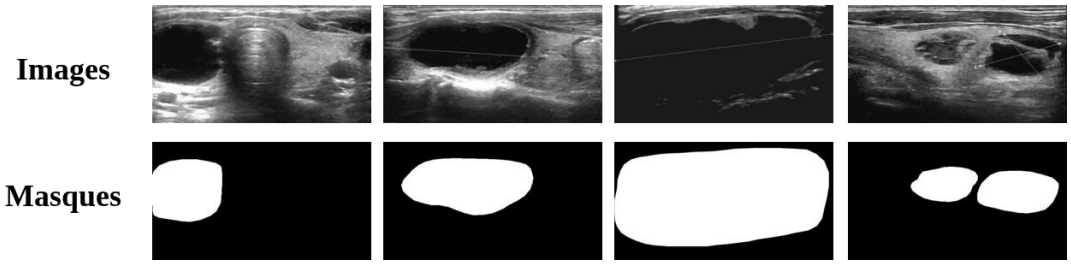


Figure 4.2 : Quelques images et masques du TN3k

Enfin, le jeu de données REFUGE [63] regroupe des images du fond d'œil utilisées pour la segmentation du disque optique et de la cup optique. Des exemples représentatifs de ces données sont présentés à la Figure 4.3.

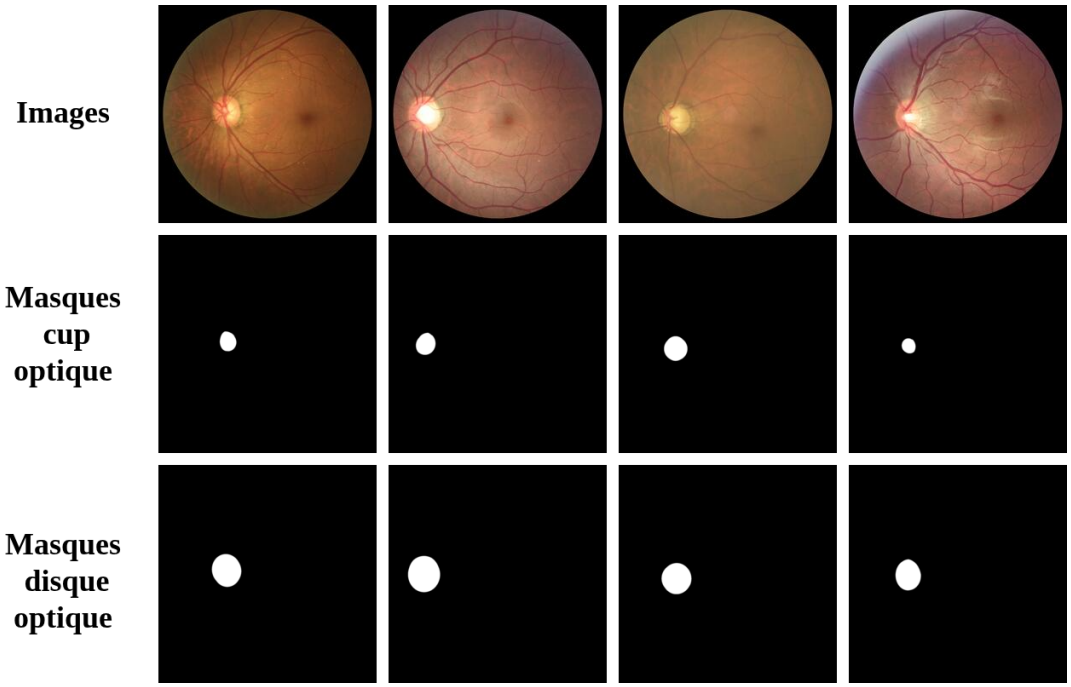


Figure 4.3 : Quelques images et masques du REFUGE

Pour chaque jeu de données, la répartition en ensembles d'entraînement et de test est celle fournie avec les données, telle que présentée dans le Tableau 4.1.

Avant les phases d'entraînement et de test, un prétraitement est appliqué pour assurer la compatibilité avec l'architecture du modèle. Les images sont redimensionnées à une résolution de 1024×1024 pixels pour correspondre aux spécifications de l'encodeur d'images de SAM. Les masques de vérité terrain sont quant à eux redimensionnés à 256×256 pixels, conformément à la résolution de sortie native du décodeur de SAM.

Jeu de données	**Entraînement**	**Test**	**Total**
ISIC	900	379	1279
TN3k	2879	614	3493
REFUGE	800	400	1200

Tableau 4.1 : Répartition des données

4.3 PARAMÈTRES D'ENTRAÎNEMENT

Les paramètres d'entraînement sont appliqués à l'ensemble des jeux de données afin de garantir une comparaison cohérente entre les différentes expériences. L'optimisation est réalisée à l'aide de l'optimiseur AdamW, avec ses paramètres par défaut : $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$ et un weight decay de 0.01. L'entraînement est mené sur 40 époques, avec un taux d'apprentissage fixé à 10^{-3}, déterminé empiriquement. La taille de lot (batch size) est fixée à 2 pour l'entraînement et à 1 pour le test, en raison des contraintes mémoire du cluster Fir.

La fonction de perte utilisée est une combinaison de la Binary Cross-Entropy (BCE) et du Dice loss. Cette combinaison est couramment adoptée en segmentation médicale en raison de sa capacité à prendre en compte à la fois la précision pixel par pixel et le recouvrement global

des régions segmentées. La perte totale est ainsi définie par

$$\mathcal{L} = \mathcal{L}_{\text{BCE}} + \mathcal{L}_{\text{Dice}}.$$

La Binary Cross-Entropy mesure l'écart entre la prédiction et la vérité terrain pour chaque pixel individuellement :

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)\right],$$

où $p_i \in [0, 1]$ est la probabilité prédite pour le pixel i, $y_i \in \{0, 1\}$ le label de vérité terrain, et N le nombre total de pixels.

La Dice loss mesure le chevauchement entre la région prédite et la région de référence :

$$\mathcal{L}_{\text{Dice}} = 1 - \frac{2\sum_i p_i y_i + s}{\sum_i p_i + \sum_i y_i + s},$$

où $\sum_i p_i y_i$ est l'intersection entre la prédiction et la vérité terrain, $\sum_i p_i + \sum_i y_i$ leur union, et $s = 10^{-6}$ un paramètre de lissage pour éviter les divisions par zéro.

4.4 MÉTRIQUES D'ÉVALUATION

Les performances du modèle proposé sont évaluées à l'aide de métriques couramment utilisées en segmentation d'images médicales : le score d'Intersection over Union (IoU) et le score de Dice. Ces mesures permettent d'évaluer quantitativement la similarité entre le masque

prédit par le modèle et le masque de vérité terrain. Elles sont particulièrement adaptées aux tâches de segmentation binaire où le déséquilibre entre la classe cible et l'arrière-plan est fréquent.

Le score IoU, également appelé indice de Jaccard, calcule le rapport entre l'intersection et l'union des régions segmentées :

$$\text{IoU} = \frac{|M_{\text{pred}} \cap M_{\text{gt}}|}{|M_{\text{pred}} \cup M_{\text{gt}}|} \tag{4.1}$$

où M_{pred} désigne le masque prédit par le modèle et M_{gt} le masque de vérité terrain.

Le coefficient de Dice est, quant à lui, utilisé pour évaluer la similarité globale. Il est défini par la relation suivante :

$$\text{Dice} = \frac{2|M_{\text{pred}} \cap M_{\text{gt}}|}{|M_{\text{pred}}| + |M_{\text{gt}}|} \tag{4.2}$$

Les résultats quantitatifs sont rapportés sous forme de moyennes calculées sur l'ensemble du jeu de test. Afin d'enrichir cette analyse, des histogrammes de distribution sont utilisés pour visualiser la dispersion des scores et évaluer la robustesse de H-SAM face à la diversité des échantillons.

Enfin, l'analyse quantitative est complétée par une étude qualitative ciblée. Pour chaque modalité d'imagerie, nous extrayons et visualisons spécifiquement les échantillons ayant obtenu les meilleurs et les pires scores de Dice. Cette sélection permet d'analyser visuellement

le comportement du modèle dans des conditions optimales ainsi que ses limites face à des cas critiques. Les figures correspondantes présentent systématiquement l'image originale, le masque de vérité terrain, la prédiction finale, les prompts fournis en entrée ainsi que les cartes d'erreur associées.

4.5 IMPLÉMENTATION LOGICIELLE

Les extraits de code présentés dans cette section détaillent les éléments fondamentaux de notre implémentation. Ils illustrent la manière dont les deux étapes de H-SAM, ainsi que les stratégies de génération automatique des prompts, ont été traduites en Python pour nos expérimentations.

4.5.1 IMPLÉMENTATION DES ÉTAPES DE H-SAM

L'adaptateur Hadamard est intégré directement au sein des blocs de l'encodeur d'images de SAM. Son implémentation, présentée à la figure 4.4, permet d'appliquer une transformation élément par élément sur les activations intermédiaires de l'encodeur.

L'adaptation Hadamard se déroule en deux phases successives. Durant la première étape, illustrée à la figure 4.5, l'encodeur d'images et l'encodeur de prompts sont gelés, tandis que seul le décodeur de masques est entraîné pour s'ajuster aux caractéristiques du domaine médical. Dans la seconde phase, présentée à la figure 4.6, seuls les paramètres de l'adaptateur et ceux de la couche de normalisation qui lui succède sont alors ajustés, ce qui garantit une mise à jour ciblée et légère des poids du modèle.

```python
import torch
import torch.nn as nn

class HadamardAdapter(nn.Module):  3 usages
    def __init__(self, proj):
        super(HadamardAdapter, self).__init__()
        self.proj = proj
        dim = proj.in_features
        self.weights = nn.Parameter(torch.ones(dim))
        self.bias = nn.Parameter(torch.zeros(dim))

    def forward(self, x):
        x = self.proj(x)
        # turn the 4d tensor into a 2d matrix
        b, c, h, w = x.shape
        x_2d = x.view(b * c * h, w)

        # multiplication
        weight_broadcasted = self.weights.expand_as(x_2d)
        x_2d = x_2d * weight_broadcasted

        # addition
        x_2d = x_2d + self.bias

        # return into the 4d shape
        x = x_2d.reshape(b, c, h, w)
        return x
```

Figure 4.4 : Implémentation de l'adaptateur Hadamard.

```python
def hadamard1(sam):  2 usages
    # ----------------------------------------------
    # IMAGE ENCODER IS FROZEN.
    # ----------------------------------------------
    for param in sam.image_encoder.parameters():
        param.requires_grad = False

    # ----------------------------------------------
    # PROMPT ENCODER IS FROZEN.
    # ----------------------------------------------
    for param in sam.prompt_encoder.parameters():
        param.requires_grad = False

    # ----------------------------------------------
    # MASK DECODER IS TRAINED.
    # ----------------------------------------------
    for param in sam.mask_decoder.parameters():
        param.requires_grad = True
```

Figure 4.5 : Implémentation de l'étape 1 de H-SAM.

```python
def hadamard2(sam):  5 usages
    # PROMPT ENCODER IS FROZEN.
    for param in sam.prompt_encoder.parameters():
        param.requires_grad = False

    # MASK DECODER IS FROZEN.
    for param in sam.mask_decoder.parameters():
        param.requires_grad = False

    # HADAMARD ADAPTER IS INSERT IN EACH BLOCK OF THE IMAGE ENCODER.
    # ALL THE IMAGE ENCODER PARAMETERS ARE FROZEN BESIDE THE HADAMARD PARAMETERS
    # AND THE NORM LAYER

    for param in sam.image_encoder.parameters():
        param.requires_grad = False

    for blk in sam.image_encoder.blocks:
        for par in blk.norm2.parameters():
            par.requires_grad = True

        # Surgery in the image encoder
        inter = blk.attn.proj
        hd = HadamardAdapter(inter)
        blk.attn.proj = hd
        blk.attn.proj.proj.requires_grad_(False)
```

Figure 4.6 : Implémentation de l'étape 2 de H-SAM.

4.5.2 IMPLÉMENTATION DES PROMPTS DE TYPE POINT

Les prompts de type point sont générés de manière automatisée à partir du masque de vérité terrain selon quatre stratégies de guidage distinctes. Ces approches incluent l'utilisation d'un point unique placé au centre de la région d'intérêt (1PT), d'un point échantillonné aléatoirement (1PTR), d'un ensemble de trois points aléatoires (3PTR), ainsi que d'une stratégie itérative pouvant générer jusqu'à trois points (3PTSI). Pour chaque scénario, les coordonnées sont initialement calculées dans l'espace du masque à une résolution de 256×256 pixels, avant d'être redimensionnées pour s'ajuster à la résolution de 1024×1024 pixels requise par l'encodeur d'images du modèle SAM.

La détermination du point central (1PT), dont l'implémentation est détaillée à la figure 4.7, repose sur le calcul du centre de masse des pixels constituant la région d'intérêt. Parallèlement, la génération de points aléatoires au sein du masque, illustrée par la figure 4.8, est employée pour produire les variantes 1PTR et 3PTR. Ce processus s'effectue par un échantillonnage de pixels appartenant à l'objet segmenté, lesquels sont ensuite convertis au format de coordonnées attendu par le modèle en leur associant systématiquement un label positif.

```python
def center_mass(mask, in_size, out_size):

    mask = torch.squeeze(mask)

    y_indices, x_indices = torch.where(mask == 1)
    x_center = torch.mean(x_indices.float())
    y_center = torch.mean(y_indices.float())

    # --- scale coordinates ---
    scale = in_size / out_size
    x_center = x_center * scale
    y_center = y_center * scale

    return {
        'point_coords': torch.tensor([[x_center, y_center]]),
        'point_labels': torch.ones(1)
    }
```

Figure 4.7 : Calcul du point central (1PT).

La stratégie de guidage la plus complexe, nommée 3PTSI, propose une simulation plus dynamique de l'interaction utilisateur en se basant sur les erreurs de segmentation successives. Ce processus débute par l'obtention d'une première prédiction à partir d'un point central initial. La comparaison entre cette prédiction et la vérité terrain permet de générer une carte d'erreur (Fig. 4.9) distinguant les zones de faux positifs et de faux négatifs.

```python
def random_click(mask, k, in_size, out_size):
    m = mask.squeeze()
    idx = torch.nonzero(m > 0)                # pixels du ROI
    pts = idx[torch.randperm(len(idx))[:k]]   # tirage aléatoire

    pts = pts[:, [1, 0]].float()              # (y,x) -> (x,y)
    pts *= in_size / out_size                 # resize 256 -> 1024

    return {
        "point_coords": pts,
        "point_labels": torch.ones(len(pts), dtype=torch.int64)
    }
```

Figure 4.8 : Génération de points aléatoires (1PTR/3PTR).

À chaque étape du cycle, la région d'erreur la plus étendue est isolée afin d'y échantillonner aléatoirement un nouveau point correctif (Fig. 4.10), lequel est ensuite agrégé aux points existants.

Ce cycle, présenté dans son ensemble à la figure 4.11, se poursuit jusqu'à l'atteinte d'un maximum de trois points ou jusqu'à la disparition des erreurs significatives, permettant ainsi une mise à jour progressive des prédictions.

4.5.3 IMPLÉMENTATION DES PROMPTS DE TYPE BOÎTE ENGLOBANTE

L'implémentation des prompts de type boîte englobante repose sur une génération automatisée visant à produire un encadrement centré sur la région d'intérêt, tout en ajustant le taux de recouvrement avec l'objet segmenté. À partir du masque binaire de référence, l'algorithme identifie la boîte minimale entourant la cible. Sa surface est ensuite modifiée pour atteindre un

```python
def compute_error_map(pred_mask, gt_mask):  1 usage

    pred_mask = (torch.sigmoid(pred_mask) > 0.5).float()
    gt_mask = gt_mask.float()

    # False Positives (FP): Predicted as 1 but actually 0
    fp = (pred_mask == 1) & (gt_mask == 0)
    # False Negatives (FN): Predicted as 0 but actually 1
    fn = (pred_mask == 0) & (gt_mask == 1)

    fp = fp.squeeze()  # -> [H,W]
    fn = fn.squeeze()

    # Error map: FP as 1 and  FN as 2
    error_map = torch.zeros_like(pred_mask, dtype=torch.uint8)
    error_map[fp] = 1
    error_map[fn] = 2

    return error_map
```

Figure 4.9 : Construction de la carte d'erreur.

```python
def largest_error_point(args, error_map):  1 usage
    em = error_map.squeeze().cpu().numpy()
    if not (em > 0).any():
        return None

    labeled, _ = label(em > 0)
    sizes = np.bincount(labeled.ravel()); sizes[0] = 0
    comp = (labeled == sizes.argmax())

    ys, xs = np.where(comp)
    i = np.random.randint(len(ys))
    y, x = ys[i], xs[i]

    lab = 0 if em[int(y), int(x)] == 1 else 1   # FP→0, FN→1
    s = args.in_size / args.out_size            # 256→1024

    return {
        "point_coords": torch.tensor( data: [[x*s, y*s]], dtype=torch.float32),
        "point_labels": torch.tensor( data: [lab], dtype=torch.int64),
    }
```

Figure 4.10 : Sélection du point correctif.

```python
def get_iterative_prompt(args, model, img, mask, device, max_pts=3):

    pts = [get_prompt(args, prompt: "1pt", model, img, mask, device)]  # (coords, labels)

    for _ in range(max_pts - 1):
        coords = torch.stack([p[0] for p in pts], dim=2).squeeze(0)
        labels = torch.stack([p[1] for p in pts], dim=2).squeeze(0)

        se, de = model.prompt_encoder((coords, labels), None, None)
        pred, _ = model.mask_decoder(
            image_embeddings=img,
            image_pe=model.prompt_encoder.get_dense_pe(),
            sparse_prompt_embeddings=se,
            dense_prompt_embeddings=de,
            multimask_output=True,
        )
        best = pred.squeeze()[find_better_mask(pred.detach().cpu(), mask)]
        error_map = compute_error_map(best.detach().cpu(), mask.detach().cpu())
        err = largest_error_point(args, error_map)
        if err is None:
            break

        pts.append((err["point_coords"].unsqueeze(0).to(device),
                    err["point_labels"].unsqueeze(0).to(device)))

    return (torch.stack([p[0] for p in pts], dim=2).squeeze(0),
            torch.stack([p[1] for p in pts], dim=2).squeeze(0))
```

Figure 4.11 : Algorithme de guidage itératif (3PTSI).

taux de recouvrement défini, ce qui permet de simuler des sélections de tailles variables pour l'expérimentation.

Une fois la boîte recentrée sur l'objet, ses coordonnées sont redimensionnées pour assurer la cohérence entre l'espace du masque (256×256 pixels) et la résolution d'entrée de l'encodeur d'images de SAM (1024×1024 pixels). Cette étape, dont l'implémentation Python est détaillée à la figure 4.12, permet l'alignement entre le prompt et les caractéristiques de l'image. Elle est utilisée pour générer les variantes de prompts nommées BBOX_0.5 et BBOX_0.75.

```python
def box_from_mask(mask, overlap, in_size=1024, out_size=256):
    m = mask.squeeze()                      # [out_size, out_size]
    ys, xs = torch.where(m > 0)             # pixels du ROI
    y0, y1 = int(ys.min()), int(ys.max()) + 1
    x0, x1 = int(xs.min()), int(xs.max()) + 1

    tw, th = x1 - x0, y1 - y0               # taille ROI
    area = (tw * th) / overlap              # aire bbox agrandie
    aspect = tw / th                        # ratio largeur/hauteur

    new_h = math.ceil(math.sqrt(area / aspect))
    new_w = math.ceil(new_h * aspect)

    cx, cy = (x0 + x1) / 2, (y0 + y1) / 2   # centre
    x0b = int(cx - new_w / 2); x1b = x0b + new_w
    y0b = int(cy - new_h / 2); y1b = y0b + new_h

    s = in_size / out_size                  # 256 -> 1024
    return [round(x0b * s), round(y0b * s), round(x1b * s), round(y1b * s)]
```

Figure 4.12 : Génération des boîtes englobantes.

4.6 ÉTUDE EXPÉRIMENTALE ET ANALYSE DES RÉSULTATS

L'étude expérimentale de ce travail s'articule autour de deux axes complémentaires. La première phase est dédiée à l'évaluation de l'impact des étapes de la stratégie Hadamard afin de sélectionner la configuration la plus performante. La seconde phase confronte cette variante aux méthodes de référence de l'état de l'art sur l'ensemble des jeux de données médicaux considérés.

4.6.1 PHASE 1 : ÉTUDE DE L'IMPACT DES ÉTAPES DE HADAMARD

Cette phase initiale, réalisée sur le jeu de données ISIC, compare la variante HSAM 1+2, intégrant les deux étapes d'adaptation, à la variante simplifiée HSAM 2, reposant exclusivement sur la seconde étape (adaptation de l'encodeur). L'objectif est de quantifier l'influence de chaque phase sur la précision finale et la stabilité de la segmentation. En complément, nous évaluons l'efficacité des diverses configurations de prompts : les points uniques (1PT, 1PTR), les points multiples (3PTS, 3PTSI) et les boîtes englobantes (BBOX_0.5, BBOX_0.75).

Prompt	**HSAM 1+2**		**HSAM 2**	
	Dice (%)	IoU (%)	Dice (%)	IoU (%)
1PT	$94,62 \pm 7,47$	$90,46 \pm 9,58$	$\mathbf{95,52 \pm 4,14}$	$\mathbf{91,68 \pm 6,38}$
1PTR	$95,10 \pm 6,16$	$91,11 \pm 7,90$	$\mathbf{95,58 \pm 4,97}$	$\mathbf{91,85 \pm 6,85}$
3PTS	$94,66 \pm 7,22$	$90,48 \pm 9,32$	$\mathbf{95,68 \pm 4,42}$	$\mathbf{91,98 \pm 6,34}$
3PTSI	$95,75 \pm 5,62$	$92,21 \pm 7,11$	$\mathbf{96,16 \pm 2,31}$	$\mathbf{92,69 \pm 4,11}$
BBOX 0.5	$95,54 \pm 5,46$	$91,78 \pm 6,58$	$\mathbf{95,96 \pm 2,79}$	$\mathbf{92,35 \pm 4,83}$
BBOX 0.75	$95,82 \pm 2,60$	$92,08 \pm 4,52$	$\mathbf{96,37 \pm 1,97}$	$\mathbf{93,05 \pm 3,56}$

Tableau 4.2 : Comparaison quantitative de la Phase 1 sur le dataset ISIC (Moyenne ± Écart-type).

L'analyse des performances rapportées dans le Tableau 4.2 montre que la variante HSAM 2 est plus performante et plus stable que la configuration complète HSAM 1+2. Sur l'ensemble des métriques, on observe une progression systématique des scores moyens : le gain sur l'IoU s'élève par exemple de $1,22$ pour le point au centre 1PT ($91,68$ % contre $90,46$ %) et atteint $1,50$ pour les points aléatoires 3PTS ($91,98$ % contre $90,48$ %). Au-delà de ces moyennes, la supériorité de cette variante réside dans la réduction drastique de la variabilité des prédictions. Alors que HSAM 1+2 montre une forte dispersion sur l'IoU du point unique au centre 1PT ($\pm 9,58$), HSAM 2 stabilise cette mesure à $\pm 6,38$. Cette tendance est encore plus marquée sur le score Dice du prompt interactif 3PTSI, où l'écart-type est divisé par plus de deux, chutant de $\pm 5,62$ à $\pm 2,41$. Cette baisse de la dispersion suggère que HSAM 2 ne se contente pas de mieux segmenter, mais le fait avec une régularité accrue face à la diversité des images.

En se concentrant sur la variante HSAM 2, on remarque une hiérarchie claire entre les types de prompts. Pour les points uniques, le mode aléatoire (1PTR) obtient une moyenne un peu plus élevée que le point central (1PT), avec $91,85 \pm 6,85$ % d'IoU contre $91,68 \pm 6,38$ %. Toutefois, le point central reste plus régulier (écart-type plus faible), ce qui en fait un repère plus fiable. Dans la catégorie des points multiples, l'interactivité (3PTSI) permet d'atteindre $92,67 \pm 4,21$ % d'IoU, marquant une nette progression par rapport aux $91,98 \pm 6,34$ % de la version statique (3PTS). Ce gain prouve que corriger le modèle réduit efficacement les erreurs. Enfin, le resserrement du cadre des boîtes englobantes à $0,75$ porte le score à son maximum de $93,05 \pm 3,56$ % d'IoU, contre $92,35 \pm 4,83$ % pour le ratio $0,5$. Ce résultat montre que plus la boîte est précise, mieux le modèle délimite la lésion. Pour la suite, nous sélectionnons

donc les trois configurations les plus représentatives : le point unique au centre 1PT, les points interactifs 3PTSI et la boîte englobante BBOX_0.75.

Pour donner une réalité visuelle à ces mesures, nous confrontons maintenant les échecs les plus marquants de nos modèles. En mettant en parallèle leurs limites respectives, nous cherchons à comprendre si une erreur de segmentation relève d'une faiblesse de l'architecture ou d'un obstacle propre à l'image, comme une pilosité abondante ou un manque de contraste. Cet examen des contours devient alors le pivot nécessaire pour valider la robustesse de l'approche gagnante HSAM 2 avant de la confronter à l'état de l'art.

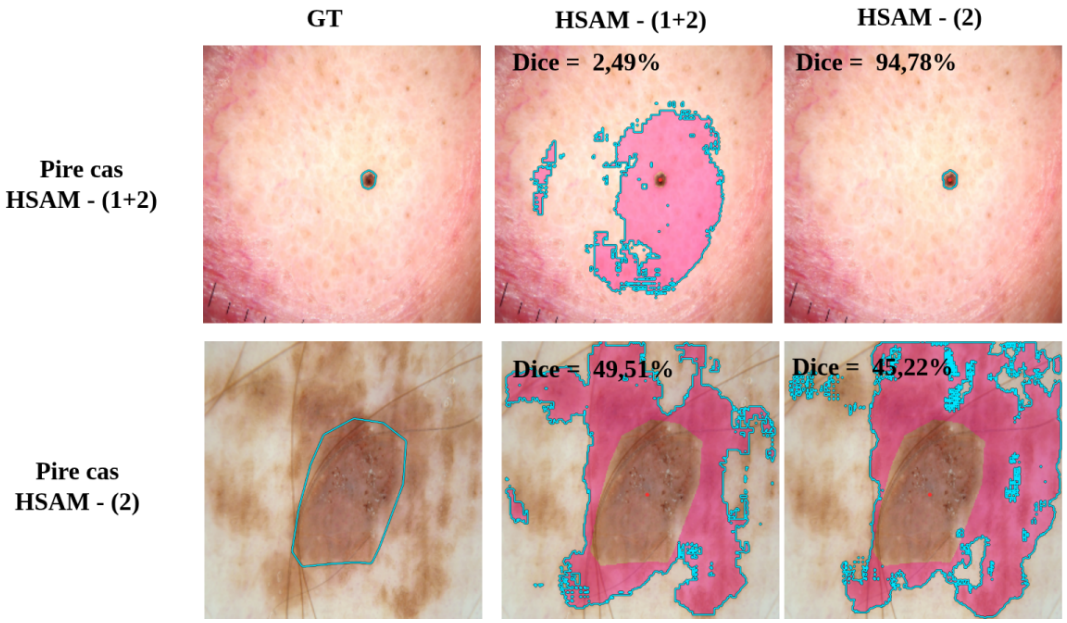


Figure 4.13 : Pires cas 1PT.

L'observation des figures 4.13, 4.14 et 4.15 montre que la variante HSAM 1+2 subit un échec structurel sur une même lésion de petite taille . Quel que soit le prompt fourni, le modèle produit une segmentation aberrante qui s'étend à la peau saine environnante. Avec un point placé au centre, le score Dice chute à $2,49\%$, tandis qu'il ne dépasse pas $5,81\%$ avec les trois

points interactifs. Même avec une boîte englobante, l'erreur demeure importante avec un score de 74, 04%. En revanche, la version HSAM 2 traite ce même échantillon avec une stabilité exemplaire, affichant des scores supérieurs à 91, 00% pour les trois types de prompts.

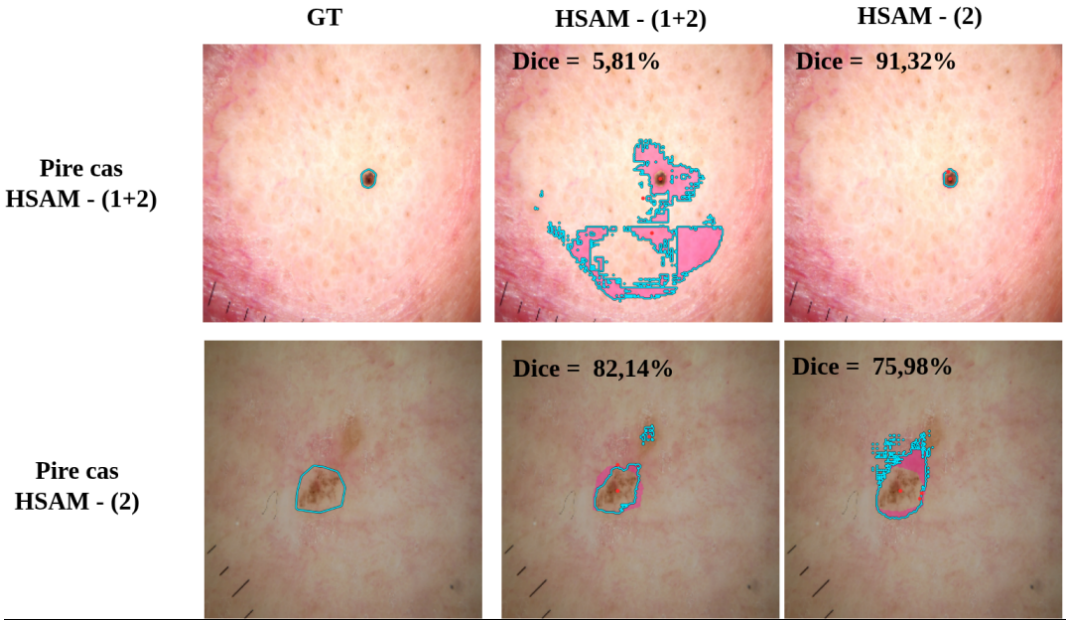


Figure 4.14 : Pires cas 3PTSI.

Les pires cas de HSAM 2 présentés dans ces figures illustrent ses limites face à une lésion caractérisée par un faible contraste entre la peau environnante et la zone à segmenter. Le score le plus bas est observé pour le point au centre avec 45, 22% de Dice. Cependant, l'apport d'information spatiale des trois points interactifs et de la boîte englobante permet de rectifier la prédiction, les scores remontant respectivement à 75, 98% et 83, 20%. Si la variante HSAM 1+2 affiche des scores supérieurs sur ces échantillons, l'écart reste toutefois marginal. L'erreur de HSAM 2 ici n'est donc pas liée à une instabilité du modèle, mais à la complexité intrinsèque de l'image.

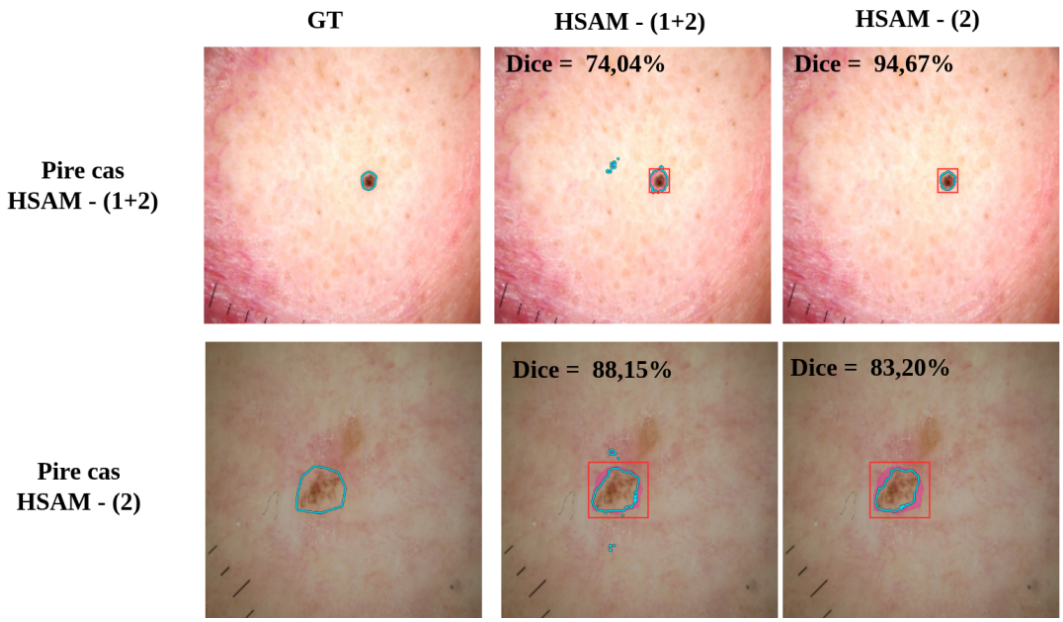


Figure 4.15 : Pires cas BBOX_075.

En conclusion, cet examen visuel valide les résultats quantitatifs globaux . L'approche HSAM 2 est privilégiée car, bien qu'elle soit légèrement moins précise sur certaines textures difficiles, elle garantit une régularité indispensable en évitant les segmentations explosives, ce qui en fait le modèle le plus fiable pour la comparaison avec les méthodes de l'état de l'art.

4.6.2 PHASE 2 : COMPARAISON AVEC LES MÉTHODES DE L'ÉTAT DE L'ART

Cette seconde phase est consacrée à la comparaison de HSAM 2 avec plusieurs méthodes de l'état de l'art. L'objectif est de situer notre approche face à SAM sans adaptation, MedSAM et Medical SAM Adapter (Med-SA), en considérant à la fois la complexité paramétrique des modèles, les performances quantitatives obtenues et l'analyse qualitative des segmentations produites. Les évaluations sont menées sur trois jeux de données couvrant des modalités variées : ISIC pour la dermoscopie, TN3K pour l'échographie thyroïdienne et REFUGE pour l'imagerie du fond d'œil. Il convient cependant de préciser que les performances relatives

à MedSAM et Med-SA ne sont pas produites dans ce travail mais sont reprises de l'article introduisant l'approche Med-SA. Comme pour l'étape précédente, l'évaluation repose sur les métriques Dice et IoU calculées sur les ensembles de test.

Complexité paramétrique des modèles

Méthode	Param. (M)
SAM	0
HSAM 2	**0,16**
Med-SA	13
MedSAM	636

Figure 4.16 : Nombre de paramètres entraînables.

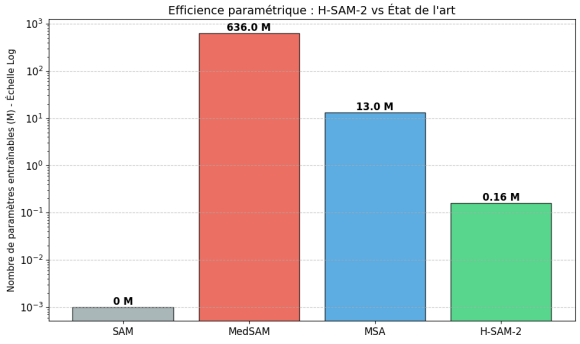


Figure 4.17 : Visualisation en échelle logarithmique.

Le tableau des paramètres 4.16 et la représentation en échelle logarithmique (Figure 4.17) mettent en évidence la faible complexité paramétrique de HSAM 2. Le modèle ne mobilise que 0,16 million de paramètres entraînables, contre 13 millions pour Med-SA et 636 millions pour MedSAM. Cette différence correspond à une réduction d'un facteur supérieur à 80 par rapport à Med-SA et d'un ordre de grandeur encore plus important par rapport à MedSAM. Malgré cette réduction significative de complexité, les résultats de segmentation restent élevés sur l'ensemble des jeux de données.

Résultats quantitatifs

Une vue d'ensemble (Figure 4.18) des performances obtenues avec le prompt de type BBOX_0.75 permet d'observer le positionnement relatif des modèles sur les trois jeux de données considérés. HSAM 2 se situe systématiquement parmi les modèles les plus performants, tout en conservant une complexité paramétrique nettement inférieure à celle des méthodes de référence. Les écarts de performance varient selon les modalités d'imagerie, mais la tendance générale montre que la réduction du nombre de paramètres n'entraîne pas de dégradation des résultats.

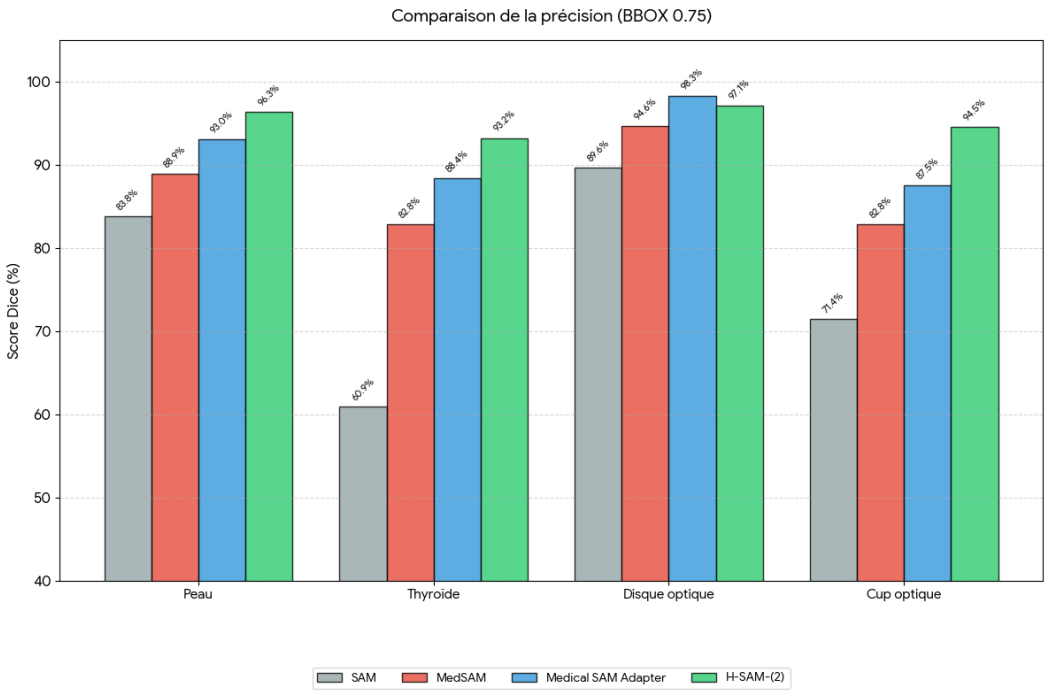


Figure 4.18 : Comparaison des résultats obtenus avec la BBOX_075.

Les performances enregistrées sur le jeu de données ISIC sont détaillées dans le tableau 4.3. Sur ce jeu de données consacré à la segmentation de lésions cutanées en dermoscopie, HSAM

2 obtient les meilleurs scores pour l'ensemble des stratégies de prompts. En configuration BBOX_0.75, le modèle atteint un Dice de 96, 3%, dépassant Med-SA d'environ trois points et MedSAM de plus de sept points. Les performances restent élevées pour les configurations à un point et à trois points, ce qui témoigne d'une bonne stabilité face aux variations de prompt. L'écart avec SAM sans adaptation demeure marqué, confirmant la nécessité d'une spécialisation du modèle pour ce type d'images.

Méthode	**Prompt**	**Dice**	**IoU**
SAM	1PT	75,6	63,7
SAM	3PTSI	77,8	66,6
SAM	BBOX 0.75	83,8	73
MedSAM	1 point	86,8	77,5
MedSAM	3 points	87,5	78,6
MedSAM	BBOX 0.75	88,9	79,8
Med-SA	1 point	92,6	84,1
Med-SA	3 points	93,4	84,7
Med-SA	BBOX 0.75	93,0	84,2
HSAM 2	1PT	95,5	91,6
HSAM 2	3PTSI	96,1	92,6
HSAM 2	BBOX 0.75	**96,3**	**93**

Tableau 4.3 : Résultats sur ISIC.

Le jeu de données TN3K, constitué d'images échographiques de nodules thyroïdiens, présente des contours plus diffus et un bruit visuel plus important. Dans ce contexte plus

difficile, HSAM 2 conserve néanmoins les meilleurs résultats. Avec le prompt BBOX_0.75, le Dice atteint 93, 23%, soit un gain d'environ cinq points par rapport à Medical SAM Adapter et un écart encore plus important vis-à-vis de SAM sans adaptation. Comme pour ISIC, les performances progressent lorsque l'information fournie par le prompt devient plus précise, ce qui confirme le rôle déterminant du guidage spatial dans ce type de segmentation.

Le tableau 4.4 regroupe les scores obtenus sur le jeu de données TN3K.

Méthode	**Prompt**	**Dice**	**IoU**
SAM	1PT	51,7	39,6
SAM	3PTSI	60,9	48,2
SAM	BBOX 0.75	73,4	60,9
MedSAM	1 point	81,3	74,7
MedSAM	3 points	81,6	75,1
MedSAM	BBOX 0.75	82,8	75,7
Med-SA	1 point	86,3	78,7
Med-SA	3 points	86,7	79,4
Med-SA	BBOX 0.75	88,4	80,4
HSAM 2	1PT	91,1	84,7
HSAM 2	3PTSI	92,7	87,2
HSAM 2	BBOX 0.75	**93,2**	**87,7**

Tableau 4.4 : Résultats sur TN3K.

Enfin, le tableau 4.5 expose les résultats sur les images du fond d'œil du jeu de données REFUGE. Ici, l'évaluation distingue les performances sur le disque optique et sur la cupule optique. HSAM 2 obtient les meilleurs scores sur la cupule optique, avec un Dice de 94, 48% en configuration BBOX_0.75. Pour le disque optique, les performances restent comparables à celles des meilleures méthodes de référence, tout en étant obtenues avec un nombre de paramètres entraînables nettement inférieur. Les gains par rapport à SAM sans adaptation sont observés pour l'ensemble des stratégies de prompts.

Méthode	**Prompt**	**Dice disc**	**IoU disc**	**Dice cup**	**IoU cup**
SAM	1PT	64,5	52,8	60	45,1
SAM	3PTSI	74,7	63,9	65,5	50,5
SAM	BBOX 0.75	89,5	81,6	71,3	56,1
MedSAM	1 point	92,9	85,5	82,1	73,8
MedSAM	3 points	93,8	86,2	82,8	74,2
MedSAM	BBOX 0.75	94,6	86,7	82.8	75,9
Med-SA	1 point	97,4	89,5	86,8	78,8
Med-SA	3 points	97,9	89,8	87,1	79,0
Med-SA	BBOX 0.75	**98.3**	90.1	87.5	79.9
HSAM 2	1PT	96,7	93,6	93,3	87,6
HSAM 2	3PTSI	97	94,3	93,6	88,1
HSAM 2	BBOX 0.75	97,1	**94,4**	**94,4**	**89,5**

Tableau 4.5 : Résultats sur REFUGE.

Dans l'ensemble, les résultats quantitatifs montrent que HSAM 2 atteint des performances supérieures ou comparables aux méthodes de l'état de l'art tout en reposant sur un nombre de paramètres entraînables très réduit. On observe également une amélioration progressive des scores lorsque l'information fournie par le prompt devient plus précise : les configurations à trois points et les boîtes englobantes avec un recouvrement de 0,75 conduisent systématiquement aux meilleurs résultats. Ces observations confirment l'efficacité de la stratégie d'adaptation proposée et son intérêt pour la segmentation d'images médicales dans un cadre paramètre-efficiente.

Analyse qualitative

Les figures 4.19, 4.21, 4.20 et 4.22 présentent des exemples de segmentations obtenues sur les différents jeux de données pour les stratégies de prompts 1PT, 3PTSI et BBOX_075. Les résultats de HSAM 2 sont comparés à ceux du modèle SAM sans adaptation. Dans ces visualisations, les points de prompts sont indiqués en rouge, les contours prédits en bleu et les régions d'erreur de segmentation en rose.

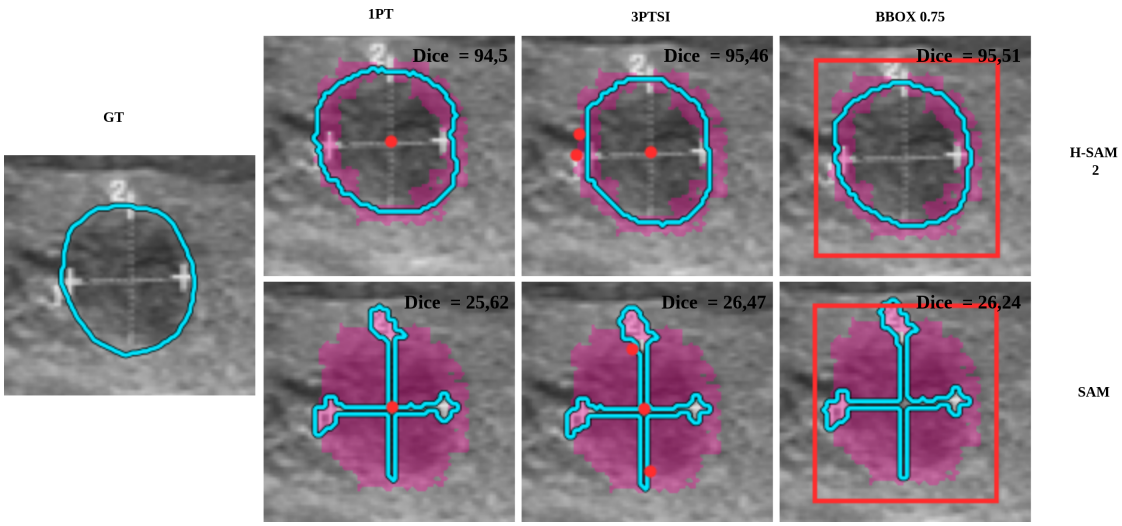


Figure 4.19 : Analyse qualitative sur le jeu de données TN3K

Sur l'exemple tn3k (figure 4.19), correspondant à une image échographique de nodule thyroïdien, l'analyse met en évidence une confusion critique du modèle SAM face aux artefacts techniques de l'appareil. Une difficulté réside ici dans la présence d'une croix de mesure blanche dont le contraste élevé occulte les bordures déjà diffuses de la lésion. Avec 1 point au centre, H-SAM 2 se distingue par sa capacité à ignorer les repères artificiels pour se concentrer sur la masse sombre du nodule. À l'inverse, SAM se laisse piéger par les lignes de la croix, échouant totalement à identifier la structure anatomique globale. L'ajout de points ne suffit pas à corriger le biais de SAM, qui reste confiné aux traits internes de l'appareil sans parvenir à s'étendre à la pathologie. H-SAM 2, en revanche, utilise cette interactivité pour lisser son contour et s'isoler des échos tissulaires environnants, sans être perturbé par le marquage central. Même sous la contrainte d'une boîte englobante, SAM persiste à ne segmenter que le repère géométrique à l'intérieur du cadre. H-SAM 2 exploite ce prompt pour verrouiller la forme globale de la lésion, prouvant sa capacité à dissocier les informations cliniques des superpositions graphiques de l'échographe.

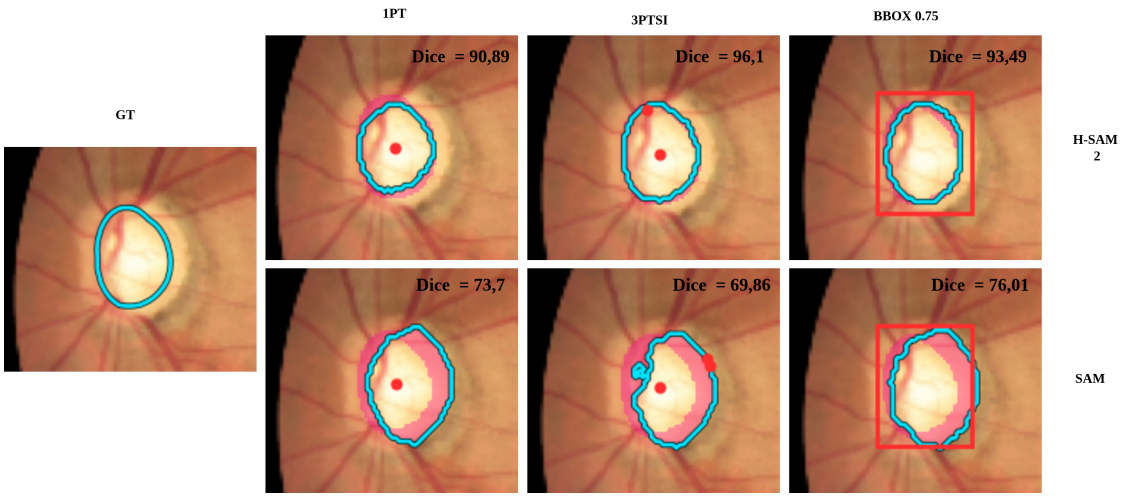


Figure 4.20 : Analyse qualitative sur la cup optique (REFUGE)

Dans le cas de la cup optique illustrée par la figure 4.20, l'enjeu majeur réside dans l'interférence des vaisseaux sanguins qui traversent des bordures déjà peu contrastées. Cette difficulté visuelle, où des lignes sombres découpent la zone claire centrale, met à l'épreuve la capacité des modèles à ne pas se laisser distraire par des éléments anatomiques extérieurs. Avec un seul point (1PT), H-SAM 2 démontre une compréhension supérieure de la structure en ignorant ces vaisseaux pour se concentrer sur la forme réelle de la cup, alors que SAM interprète mal ces limites et produit une sursegmentation marquée. L'ajout de points (3PTSI) accentue cette divergence : SAM devient instable, produisant un contour irrégulier en "dents de scie" qui tente maladroitement de contourner les vaisseaux, tandis que H-SAM 2 conserve une courbe fluide et organique fidèle à la morphologie réelle. Enfin, bien que la boîte englobante (BBOX 0.75) aide à recentrer les résultats, SAM continue de surestimer la partie inférieure de la structure. H-SAM 2, au contraire, utilise cette contrainte pour affiner ses bords et s'ajuster avec précision à la vérité terrain, prouvant sa robustesse face au bruit anatomique complexe du fond d'œil.

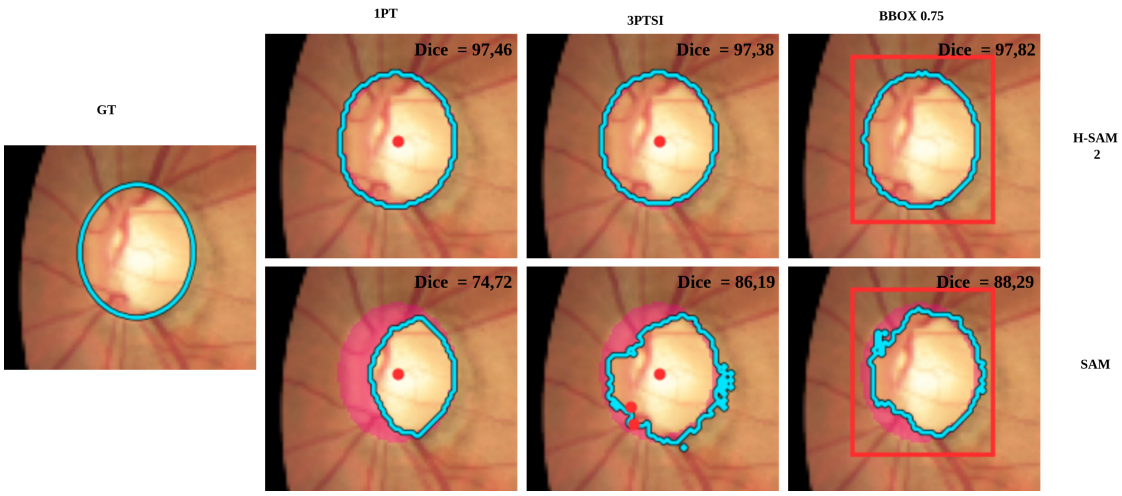


Figure 4.21 : Analyse qualitative sur le disque optique (REFUGE)

Pour le disque optique (figure 4.21), la difficulté réside dans l'étendue de la structure et l'interférence des vaisseaux qui masquent les contours périphériques. Ce cliché se caractérise par une zone d'intérêt plus large que la cup, nécessitant une capacité à capturer la globalité de l'organe malgré les obstacles vasculaires. Avec un point au centre, H-SAM 2 démontre une efficacité immédiate en appréhendant la quasi-totalité de la forme dès le premier prompt, tandis que SAM échoue à percevoir les limites réelles et produit une segmentation très incomplète, restreinte au noyau central. L'ajout de points met en lumière une fragilité structurelle chez SAM : bien que le masque s'élargisse, il devient extrêmement irrégulier, présentant des bordures hachées qui tentent maladroitement de suivre les contours. H-SAM 2, à l'inverse, conserve un tracé fluide et organique, prouvant sa stabilité face à la complexité des tissus. Sous la contrainte d'une boîte englobante, H-SAM 2 verrouille la forme avec une précision rigoureuse. SAM, bien que mieux cadré, conserve des irrégularités périphériques et ne parvient pas à égaler la netteté du contour de son concurrent, qui reste le plus fidèle à la vérité terrain.

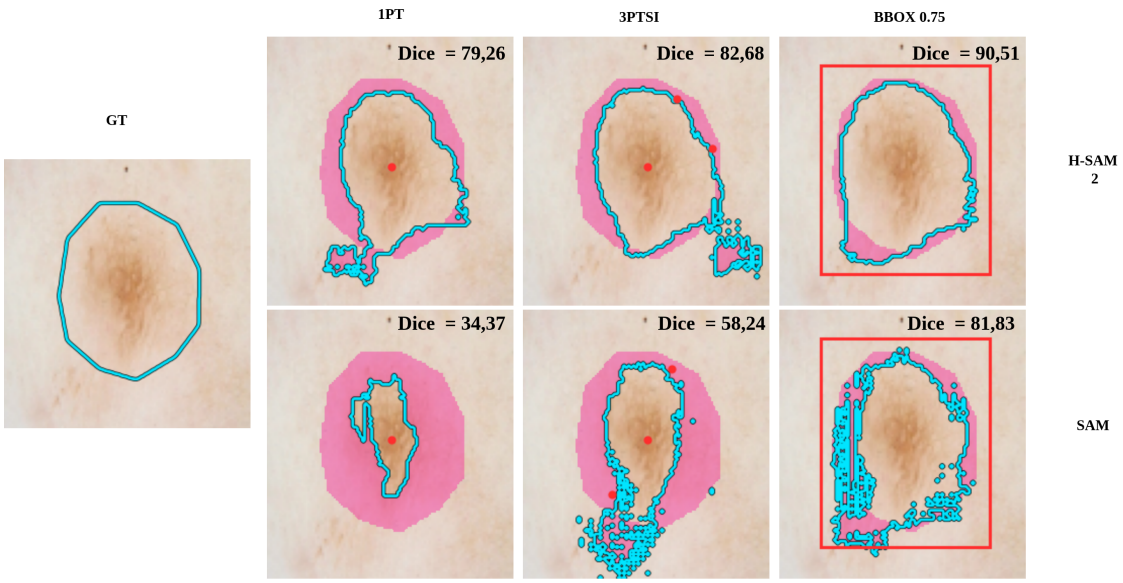


Figure 4.22 : Analyse qualitative sur le jeu de données ISIC

Enfin, la figure 4.22 expose une lésion dermoscopique caractérisée par une transition chromatique diffuse entre le pigment brun et la peau saine. Ce dégradé efface les limites franches, rendant la détection des marges extrêmement vulnérable aux variations de texture et de teinte. Avec un seul point (1PT), H-SAM 2 appréhende d'emblée l'envergure globale de la masse, alors que SAM s'isole sur un noyau central restreint, incapable d'interpréter les zones de faible contraste périphériques. L'ajout de points interactifs (3PTSI) met en évidence une instabilité morphologique chez SAM, qui produit des amas de pixels fragmentés et des bordures « bruitées », tandis que H-SAM 2 maintient une enveloppe fluide et organique. Enfin, sous la contrainte d'une boîte englobante (BBOX 0.75), H-SAM 2 génère un contour continu s'ajustant avec précision au dégradé pigmentaire. SAM, au contraire, présente des hachures massives et des irrégularités sur ses flancs, témoignant d'une difficulté persistante à stabiliser son masque par rapport à la vérité terrain.

Dans l'ensemble, ces visualisations montrent que la qualité de la segmentation dépend fortement de l'information spatiale fournie par les prompts. Lorsque les contours sont complexes ou peu contrastés, l'ajout de points ou l'utilisation de boîtes englobantes permet de réduire les régions d'erreur et d'améliorer l'ajustement des contours. À prompt équivalent, H-SAM 2 produit généralement des segmentations plus compactes et plus stables que SAM, avec des erreurs moins étendues et des contours mieux alignés avec la vérité terrain. Ces observations visuelles complètent les résultats quantitatifs présentés précédemment et permettent d'illustrer concrètement le comportement des modèles sur différents types d'images.

4.7 DISCUSSIONS

Les résultats obtenus dans cette étude mettent en évidence l'efficacité d'une stratégie d'adaptation paramètre-efficiente du modèle SAM pour la segmentation d'images médicales. Malgré un nombre de paramètres entraînables très réduit, HSAM 2 atteint des performances supérieures ou comparables aux méthodes de référence sur l'ensemble des jeux de données considérés. Ce constat confirme que l'adaptation ciblée des représentations internes du modèle, via l'introduction d'adaptateurs Hadamard, permet de réduire le fossé sémantique entre les images naturelles utilisées lors du pré-entraînement de SAM et les structures anatomiques présentes dans les images médicales. L'analyse quantitative montre que les gains sont particulièrement marqués sur les modalités où les contours sont peu contrastés, comme l'échographie thyroïdienne. Dans ce contexte, la spécialisation du décodeur puis la modulation progressive des représentations internes contribuent à stabiliser la prédiction et à limiter les erreurs périphériques. Sur des structures plus régulières, comme le disque optique, les

performances restent comparables aux meilleures méthodes existantes, tout en étant obtenues avec un coût paramétrique nettement inférieur. Ces observations suggèrent que la stratégie proposée constitue un compromis efficace entre précision de segmentation et légèreté du modèle. Les résultats mettent également en évidence l'importance du guidage spatial fourni par les prompts. Les configurations utilisant plusieurs points ou des boîtes englobantes précises conduisent systématiquement aux meilleures performances, ce qui souligne la dépendance des modèles interactifs à la qualité de l'information fournie en entrée. Toutefois, même dans les configurations les plus simples, HSAM 2 conserve une segmentation plus stable que SAM sans adaptation, ce qui indique que l'adaptation proposée améliore la robustesse du modèle face à des prompts moins informatifs. Certaines limites doivent néanmoins être soulignées. Les performances de HSAM 2 restent conditionnées par la précision des prompts et par la qualité des annotations de référence. Par ailleurs, les comparaisons avec MedSAM et Med-SA reposent sur des résultats issus de la littérature et non sur une reproduction expérimentale complète dans le cadre de ce travail. Enfin, bien que les résultats obtenus sur plusieurs modalités d'imagerie soient encourageants, l'évaluation pourrait être étendue à d'autres organes et à d'autres types de données afin de confirmer la capacité de généralisation de la méthode.

Dans l'ensemble, cette étude montre qu'une adaptation légère et progressive de SAM peut permettre d'atteindre des performances élevées en segmentation médicale sans recourir à un fine-tuning complet du modèle. L'approche HSAM 2 ouvre ainsi des perspectives pour le développement de méthodes d'adaptation paramètre-efficiente applicables à d'autres tâches d'imagerie médicale ou à d'autres architectures de modèles fondation.

CHAPITRE V

CONCLUSION

La segmentation d'images médicales constitue une étape fondamentale de l'analyse clinique, mais son automatisation reste complexe face à la grande diversité des modalités d'imagerie et à la variabilité anatomique des patients. Historiquement, cette tâche a reposé sur des réseaux spécialisés dont la capacité de généralisation est souvent limitée par le manque de données annotées. L'émergence du modèle Segment Anything (SAM) a marqué un tournant majeur en introduisant une approche universelle guidée par des instructions. Toutefois, comme nous l'avons soulevé en début de recherche, l'application directe de ce modèle généraliste au domaine médical rencontre des limites de précision, ce qui impose le développement de stratégies d'adaptation performantes mais économes.

Pour répondre à cet enjeu, nous avons proposé H-SAM-(2), une architecture s'appuyant sur l'intégration de modules d'adaptation Hadamard au sein de l'encodeur d'images. Cette solution a été évaluée à travers une série d'expériences couvrant trois domaines distincts : la dermatologie, l'échographie thyroïdienne et l'imagerie du fond de l'œil. L'analyse des résultats a montré que la performance du modèle est intimement liée à la nature des instructions fournies. Si la configuration BBOX 0.75 a permis d'atteindre les scores les plus élevés, notamment un

score de Dice de 96, 3% sur les lésions cutanées, les tests effectués avec plusieurs configurations des prompts ont également révélé une capacité d'adaptation intéressante, bien que plus sensible à la précision du placement manuel. Cette progression des scores en fonction de la qualité des prompts confirme que notre architecture parvient à capturer efficacement l'information spatiale pour affiner ses prédictions.

La contribution majeure de cette recherche réside dans la démonstration qu'une spécialisation fine du modèle peut être accomplie sans sacrifier la légèreté technique. En n'utilisant que 0, 16 millions de paramètres entraînables, H-SAM-(2) parvient à rivaliser avec des architectures beaucoup plus denses comme MedSAM ou Medical SAM Adapter. Nous avons ainsi prouvé qu'il est possible de réduire le coût d'adaptation par un facteur considérable tout en maintenant une précision compétitive. Cette efficacité paramétrique constitue une avancée vers des modèles plus sobres, capables d'être spécialisés sur de petits jeux de données sans risquer le surapprentissage, répondant ainsi directement au problème de la rareté des données annotées mentionné dans notre problématique.

Malgré ces performances, l'étude a mis en lumière certaines limites qu'il convient de souligner. L'analyse des résultats sur le disque optique a révélé que pour des formes géométriques simples et très régulières, les modèles de référence plus denses conservent un léger avantage sur le score de Dice. Notre méthode semble également présenter une certaine dépendance à la qualité initiale du prompt ; une instruction imprécise ou un point mal positionné peut entraîner une segmentation sous-optimale sur des structures très fines. De plus, l'évaluation s'est limitée à des images en deux dimensions, ce qui restreint pour l'instant l'application de H-SAM-(2) à des examens ne nécessitant pas de reconstruction volumétrique complexe. Enfin, la sensibilité

du modèle au bruit visuel, particulièrement en échographie, rappelle que l'adaptation légère peut parfois manquer de la robustesse nécessaire face à des images de très basse qualité.

Ces constatations ouvrent des perspectives de recherche stimulantes pour l'avenir. Une évolution naturelle de ce travail consisterait à porter l'architecture vers des modalités en trois dimensions, telles que l'imagerie par résonance magnétique (IRM) ou la tomographie, afin de vérifier si la légèreté des modules Hadamard peut être maintenue lors du traitement de volumes. Il serait également pertinent d'explorer l'automatisation de la génération des prompts pour réduire la dépendance à l'intervention humaine, par exemple en couplant notre modèle à un détecteur d'objets léger. Une autre piste résiderait dans l'hybridation des adaptateurs, en combinant les modules Hadamard avec des techniques de recalage spatial pour améliorer la précision sur les structures à faible contraste. Ce mémoire ne constitue donc pas un point final, mais propose une base pour le développement d'outils d'aide au diagnostic plus agiles et accessibles, capables de concilier la puissance des modèles de fondation avec les exigences de précision du monde médical.

BIBLIOGRAPHIE

[1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv :2303.08774*, 2023.

[2] F. F. Alqahtani, "Spect/ct and pet/ct, related radiopharmaceuticals, and areas of application and comparison," *Saudi Pharmaceutical Journal*, vol. 31, n°. 2, p. 312–328, 2023.

[3] S. Asgari Taghanaki, K. Abhishek, J. P. Cohen, J. Cohen-Adad, et G. Hamarneh, "Deep semantic segmentation of natural and medical images : a review," *Artificial intelligence review*, vol. 54, n°. 1, p. 137–178, 2021.

[4] I. Bickle, "Radiographies médicales (rid : 33325, 37964)," Radiopaedia.org, 2026, accédé le 25 mai 2026. Licence CC BY-NC-SA. [En ligne]. Disponible : https://radiopaedia.org/cases/33325

[5] R. Bommasani *et al.*, "On the opportunities and risks of foundation models," *arXiv :2108.07258*, 2021.

[6] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, p. 1877–1901, 2020.

[7] M. Caron *et al.*, "Emerging properties in self-supervised vision transformers," dans *ICCV*, 2021.

[8] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, et Y. Zhou, "Transunet : Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv :2102.04306*, 2021.

[9] X. Chen, "Real-time semantic segmentation algorithms for enhanced augmented reality," *Journal of Computational Innovation*, vol. 3, n°. 1, 2023.

[10] Y. Chen, Q. Fu, G. Fan, L. Du, J.-G. Lou, S. Han, D. Zhang, Z. Li, et Y. Xiao, “Hadamard adapter : An extreme parameter-efficient adapter tuning method for pre-trained language models,” dans *Proceedings of the 32nd ACM international conference on information and knowledge management*, 2023, p. 276–285.

[11] D. Cheng, Z. Qin, Z. Jiang, S. Zhang, Q. Lao, et K. Li, “Sam on medical images : A comprehensive study on three prompt modes,” *arXiv preprint arXiv :2305.00035*, 2023.

[12] S. H. Choi, K.-B. Park, D. H. Roh, J. Y. Lee, M. Mohammed, Y. Ghasemi, et H. Jeong, “An integrated mixed reality system for safety-aware human-robot collaboration using deep learning and digital twin generation,” *Robotics and Computer-Integrated Manufacturing*, vol. 73, p. 102258, 2022.

[13] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, et O. Ronneberger, “3d u-net : learning dense volumetric segmentation from sparse annotation,” dans *International conference on medical image computing and computer-assisted intervention*. Springer, 2016, p. 424–432.

[14] P.-H. Conze, G. Andrade-Miranda, V. K. Singh, V. Jaouen, et D. Visvikis, “Current and emerging trends in medical image segmentation with deep learning,” *IEEE Transactions on Radiation and Plasma Medical Sciences*, vol. 7, n^{o}. 6, p. 545–569, 2023.

[15] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, et L. Fei-Fei, “Imagenet : A large-scale hierarchical image database,” dans *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, p. 248–255.

[16] R. Deng, C. Cui, Q. Liu, T. Yao, L. W. Remedios, S. Bao, B. A. Landman, L. E. Wheless, L. A. Coburn, K. T. Wilson *et al.*, “Segment anything model (sam) for digital pathology : Assess zero-shot segmentation on whole slide imaging,” dans *IS&T International Symposium on Electronic Imaging*, vol. 37, 2025, p. COIMG–132.

[17] J. Donahue *et al.*, “Decaf : A deep convolutional activation feature for generic visual recognition,” dans *ICML*, 2014.

[18] A. Dosovitskiy, “An image is worth 16x16 words : Transformers for image recognition at scale,” *arXiv preprint arXiv :2010.11929*, 2020.

[19] W. Feng, L. Zhu, et L. Yu, “Cheap lunch for medical image segmentation by fine-tuning sam on few exemplars,” dans *International MICCAI Brainlesion Workshop*. Springer, 2023, p. 13–22.

[20] K. Fukushima, “Neocognitron : A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position,” *Biological cybernetics*, vol. 36, n^{o}. 4, p. 193–202, 1980.

[21] Y. Ganin et V. Lempitsky, “Unsupervised domain adaptation by backpropagation,” dans *ICML*, 2015.

[22] A. Gilbert, M. Bonny, et H. Arisen, “A systematic review of the imaging modalities used for image acquisition.” 2024.

[23] H. Gong, G. Chen, R. Wang, X. Xie, M. Mao, Y. Yu, F. Chen, et G. Li, “Multi-task learning for thyroid nodule segmentation with thyroid region prior,” dans *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2021, p. 257–261.

[24] D. Gutman, N. C. Codella, E. Celebi, B. Helba, M. Marchetti, N. Mishra, et A. Halpern, “Skin lesion analysis toward melanoma detection : Isic 2016 challenge,” *arXiv preprint arXiv :1605.01397*, 2016.

[25] I. R. I. Haque et J. Neubert, “Deep learning approaches to biomedical image segmentation,” *Informatics in Medicine Unlocked*, vol. 18, p. 100297, 2020.

[26] K. He, X. Zhang, S. Ren, et J. Sun, “Deep residual learning for image recognition,” dans *CVPR*, 2016.

[27] K. He, H. Fan, Y. Wu, S. Xie, et R. Girshick, “Momentum contrast for unsupervised visual representation learning,” dans *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[28] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, et R. Girshick, “Masked autoencoders are scalable vision learners,” dans *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[29] S. He, R. Bao, J. Li, P. E. Grant, et Y. Ou, “Accuracy of segment-anything model (sam) in medical image segmentation tasks,” *CoRR*, 2023.

[30] S. He, R. Bao, J. Li, J. Stout, A. Bjørnerud, P. E. Grant, et Y. Ou, “Computer-vision benchmark segment-anything model (sam) in medical images : Accuracy in 12 datasets,” *arXiv preprint arXiv :2304.09324*, 2023.

[31] M. H. Hesamian, W. Jia, X. He, et P. Kennedy, “Deep learning techniques for medical image segmentation : achievements and challenges,” *Journal of digital imaging*, vol. 32, n^{o}. 4, p. 582–596, 2019.

[32] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, et S. Gelly, "Parameter-efficient transfer learning for nlp," dans *International conference on machine learning*. PMLR, 2019, p. 2790–2799.

[33] N. Houlsby *et al.*, "Parameter-efficient transfer learning for nlp," dans *ICML*, 2019.

[34] E. Hu *et al.*, "Lora : Low-rank adaptation of large language models," *arXiv :2106.09685*, 2022.

[35] M. Hu, Y. Li, et X. Yang, "Skinsam : Empowering skin cancer segmentation with segment anything model," *arXiv preprint arXiv :2304.13973*, 2023.

[36] Y. Huang, X. Yang, L. Liu, H. Zhou, A. Chang, X. Zhou, R. Chen, J. Yu, J. Chen, C. Chen *et al.*, "Segment anything model for medical images ?" *Medical Image Analysis*, vol. 92, p. 103061, 2024.

[37] D. H. Hubel et T. N. Wiesel, "Receptive fields and functional architecture of monkey striate cortex," *The Journal of physiology*, vol. 195, n°. 1, p. 215–243, 1968.

[38] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, et K. H. Maier-Hein, "nnu-net : a self-configuring method for deep learning-based biomedical image segmentation," *Nature methods*, vol. 18, n°. 2, p. 203–211, 2021.

[39] S. M. S. Islam, M. A. A. Nasim, I. Hossain, D. M. A. Ullah, D. K. D. Gupta, et M. M. H. Bhuiyan, "Introduction of medical imaging modalities," dans *Data driven approaches on medical imaging*. Springer, 2023, p. 1–25.

[40] G.-P. Ji, D.-P. Fan, P. Xu, M.-M. Cheng, B. Zhou, et L. Van Gool, "Sam struggles in concealed scenes–empirical study on segment anything," *arXiv preprint arXiv :2304.06022*, 2023.

[41] H. Kasban, M. El-Bendary, D. Salama *et al.*, "A comparative study of medical imaging techniques," *International Journal of Information Science and Intelligent System*, vol. 4, n°. 2, p. 37–58, 2015.

[42] S. Khazem, "Topolora-sam : Topology-aware parameter-efficient adaptation of foundation segmenters for thin-structure and cross-domain binary semantic segmentation," *arXiv preprint arXiv :2601.02273*, 2026.

[43] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," dans *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, p. 4015–4026.

[44] A. Krizhevsky, I. Sutskever, et G. Hinton, “Imagenet classification with deep convolutional neural networks,” dans *NeurIPS*, 2012.

[45] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard, et L. Jackel, “Handwritten digit recognition with a back-propagation network,” *Advances in neural information processing systems*, vol. 2, 1989.

[46] Y. LeCun, L. Bottou, Y. Bengio, et P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, n°. 11, p. 2278–2324, 2002.

[47] L. Lei, Q. Yang, L. Yang, T. Shen, R. Wang, et C. Fu, “Deep learning implementation of image segmentation in agricultural applications : a comprehensive review,” *Artificial Intelligence Review*, vol. 57, n°. 6, p. 149, 2024.

[48] P. Leong, “Radiographie médicale (rid : 26618),” Radiopaedia.org, 2026, accédé le 25 mai 2026. Licence CC BY-NC-SA. [En ligne]. Disponible : https://radiopaedia.org/cases/26618

[49] X. Li et P. Liang, “Prefix-tuning : Optimizing continuous prompts for generation,” *ACL*, 2021.

[50] X. Li, H. Chen, X. Qi, Q. Dou, C.-W. Fu, et P.-A. Heng, “H-denseunet : hybrid densely connected unet for liver and tumor segmentation from ct volumes,” *IEEE transactions on medical imaging*, vol. 37, n°. 12, p. 2663–2674, 2018.

[51] Y. Li, M. Hu, et X. Yang, “Polyp-sam : Transfer sam for polyp segmentation,” dans *Medical imaging 2024 : computer-aided diagnosis*, vol. 12927. SPIE, 2024, p. 749–754.

[52] H. Liu *et al.*, “Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning,” *NeurIPS*, 2022.

[53] X. Liu, L. Song, S. Liu, et Y. Zhang, “A review of deep-learning-based medical image segmentation methods,” *Sustainability*, vol. 13, n°. 3, p. 1224, 2021.

[54] M. Long *et al.*, “Learning transferable features with deep adaptation networks,” dans *ICML*, 2015.

[55] ——, “Deep transfer learning with joint adaptation networks,” dans *ICML*, 2017.

[56] J. Ma, Y. He, F. Li, L. Han, C. You, et B. Wang, “Segment anything in medical images,” *Nature Communications*, vol. 15, n°. 1, p. 654, 2024.

[57] J. Maitre, K. Bouchard, et L. P. Bédard, “Mineral grains recognition using computer vision and machine learning,” *Computers & Geosciences*, vol. 130, p. 84–93, 2019.

[58] M. A. Mazurowski, H. Dong, H. Gu, J. Yang, N. Konz, et Y. Zhang, “Segment anything model for medical image analysis : an experimental study,” *Medical Image Analysis*, vol. 89, p. 102918, 2023.

[59] N. McConnell, N. Ndipenoch, Y. Cao, A. Miron, et Y. Li, “Exploring advanced architectural variations of nnunet,” *Neurocomputing*, vol. 560, p. 126837, 2023.

[60] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, et D. Terzopoulos, “Image segmentation using deep learning : A survey,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, n°. 7, p. 3523–3542, 2021.

[61] O. Oktay, J. Schlemper, L. Le Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Hammerla, B. Kainz *et al.*, “Attention u-net : Learning where to look for the pancreas,” dans *arXiv preprint arXiv :1804.03999*, 2018.

[62] M. Oquab *et al.*, “Dinov2 : Learning robust visual features without supervision,” *arXiv :2304.07193*, 2023.

[63] J. I. Orlando *et al.*, “Refuge challenge : A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs,” *arXiv preprint arXiv :1910.03667*, 2019.

[64] S. J. Pan et Q. Yang, “A survey on transfer learning,” *IEEE Transactions on knowledge and data engineering*, vol. 22, n°. 10, p. 1345–1359, 2009.

[65] D. L. Pham, C. Xu, et J. L. Prince, “Current methods in medical image segmentation,” *Annual review of biomedical engineering*, vol. 2, n°. 1, p. 315–337, 2000.

[66] I. Qureshi, J. Yan, Q. Abbas, K. Shaheed, A. B. Riaz, A. Wahid, M. W. J. Khan, et P. Szczuko, “Medical image segmentation using deep semantic-based methods : A review of techniques, applications and emerging trends,” *Information Fusion*, vol. 90, p. 316–352, 2023.

[67] A. Radford *et al.*, “Learning transferable visual models from natural language supervision,” dans *ICML*, 2021.

[68] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, et I. Sutskever, “Zero-shot text-to-image generation,” dans *International conference on machine learning*. Pmlr, 2021, p. 8821–8831.

[69] B. Rasuli, “Radiographie médicale (rid : 154347),” Radiopaedia.org, 2026, accédé le 25 mai 2026. Licence CC BY-NC-SA. [En ligne]. Disponible : https://radiopaedia.org/cases/154347

[70] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, et B. Ommer, “High-resolution image synthesis with latent diffusion models,” dans *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, p. 10 684–10 695.

[71] O. Ronneberger, P. Fischer, et T. Brox, “U-net : Convolutional networks for biomedical image segmentation,” dans *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, p. 234–241.

[72] S. Roy, T. Wald, G. Koehler, M. R. Rokuss, N. Disch, J. Holzschuh, D. Zimmerer, et K. H. Maier-Hein, “Sam. md : Zero-shot medical image segmentation capabilities of the segment anything model,” *arXiv preprint arXiv :2304.05396*, 2023.

[73] A. Sharif Razavian, H. Azizpour, J. Sullivan, et S. Carlsson, “Cnn features off-the-shelf : An astounding baseline for recognition,” dans *CVPR Workshops*, 2014.

[74] P. Shi, J. Qiu, S. M. D. Abaxi, H. Wei, F. P.-W. Lo, et W. Yuan, “Generalist vision foundation models for medical imaging : A case study of segment anything model on zero-shot medical segmentation,” *Diagnostics*, vol. 13, n°. 11, p. 1947, 2023.

[75] W. Shi, P. Zhang, Y. Li, et Z. Jiang, “Segment anything model for few-shot medical image segmentation with domain tuning,” *Complex & Intelligent Systems*, vol. 11, n°. 1, p. 37, 2025.

[76] L. Tang, H. Xiao, et B. Li, “Can sam segment anything ? when sam meets camouflaged object detection,” *arXiv preprint arXiv :2304.04709*, 2023.

[77] E. Tzeng, J. Hoffman, K. Saenko, et T. Darrell, “Adversarial discriminative domain adaptation,” dans *CVPR*, 2017.

[78] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, et I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.

[79] R. Wang, T. Lei, R. Cui, B. Zhang, H. Meng, et A. K. Nandi, “Medical image segmentation using deep learning : A survey,” *IET image processing*, vol. 16, n°. 5, p. 1243–1267, 2022.

[80] J. Wu, Z. Wang, M. Hong, W. Ji, H. Fu, Y. Xu, M. Xu, et Y. Jin, “Medical sam adapter : Adapting segment anything model for medical image segmentation,” *Medical image analysis*, vol. 102, p. 103547, 2025.

[81] X. Xiao, S. Lian, Z. Luo, et S. Li, “Weighted res-unet for high-quality retina vessel segmentation,” dans *2018 9th international conference on information technology in medicine and education (ITME)*. IEEE, 2018, p. 327–331.

[82] J. Yosinski, J. Clune, Y. Bengio, et H. Lipson, “How transferable are features in deep neural networks ?” dans *NeurIPS*, 2014.

[83] K. Zhang et D. Liu, “Customized segment anything model for medical image segmentation,” *arXiv preprint arXiv :2304.13785*, 2023.

[84] Z. Zhong, Z. Tang, T. He, H. Fang, et C. Yuan, “Convolution meets lora : Parameter efficient finetuning for segment anything model,” *arXiv preprint arXiv :2401.17868*, 2024.

[85] H.-Y. Zhou, J. Guo, Y. Zhang, L. Yu, L. Wang, et Y. Yu, “nnformer : Interleaved transformer for volumetric segmentation,” *arXiv preprint arXiv :2109.03201*, 2021.

[86] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, et J. Liang, “Unet++ : A nested u-net architecture for medical image segmentation,” dans *International workshop on deep learning in medical image analysis*. Springer, 2018, p. 3–11.