UNIVERSITÀ
DEGLI STUDI
DI PADOVA

DIPARTIMENTO
DI INGEGNERIA
DELL'INFORMAZIONE

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

CORSO DI LAUREA IN INGEGNERIA INFORMATICA

"SEGMENTAZIONE SEMANTICA: PROMPT ENGINEERING PER IL MODELLO SEGMENT ANYTHING"

Relatore: Professore Carlo Fantozzi

Laureando: Michele Veneziani

ANNO ACCADEMICO 2023-2024

Data di laurea 19/11/2024

Indice

Introduzione

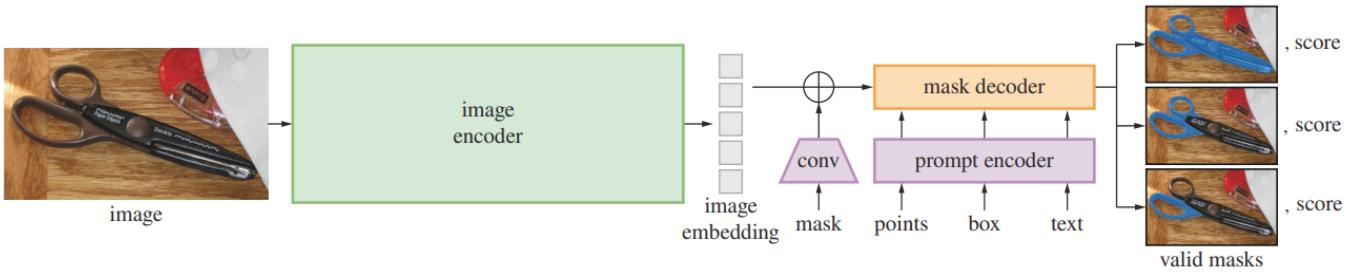
La segmentazione di immagini è un processo fondamentale nella visione artificiale che consiste nel suddividere un'immagine in regioni significative, assegnando a ciascuna di esse un'etichetta semantica (es. persona, auto, edificio). In altre parole, mira a comprendere cosa c'è nell'immagine, andando oltre la semplice individuazione dei contorni. La segmentazione semantica è una sottocategoria di questa attività, che si concentra proprio sull'assegnazione di etichette semantiche a ciascun pixel. In questo contesto, i modelli di segmentazione rappresentano lo State Of The Art, ovvero la frontiera della ricerca in un determinato campo (le tecniche e gli algoritmi più recenti e performanti).

Questi modelli, basati su tecniche di deep learning, hanno compiuto progressi significativi negli ultimi anni, consentendo di ottenere risultati sempre più accurati e robusti. Tra questi, il Segment Anything Model (SAM) si distingue per le sue capacità innovative. SAM, in quanto foundation model (un modello di intelligenza artificiale pre-addestrato su un'enorme quantità di dati, che può essere adattato a diversi compiti di segmentazione senza dover essere riaddestrato completamente), si basa su un approccio generativo, producendo output (maschere di segmentazione) a partire da input testuali o puntatori. La valutazione della performance di tali modelli richiede l'utilizzo di metriche specifiche come il Dice Similarity Coefficient e la Hausdorff Distance, che quantificano rispettivamente la sovrapposizione e la distanza tra le segmentazioni predette e quelle ground truth. Il Jaccard Similarity Index, noto anche come IoU, fornisce un'ulteriore misura della performance, in particolare è definito come il rapporto tra l'area di sovrapposizione tra la segmentazione prevista e la segmentazione di ground truth e l'area di unione tra la segmentazione prevista e la segmentazione di ground truth, particolarmente utile per valutare la precisione delle segmentazioni. L'architettura di SAM si basa sui Vision Transformer (ViT), in particolare sulle varianti ViT-B (Base) e ViT-H (Huge). Questi modelli, che differiscono per dimensione e complessità, hanno dimostrato di essere particolarmente efficaci per compiti di segmentazione.

In questa tesi, ci proponiamo di analizzare in profondità il Segment Anything Model, esplorandone le caratteristiche distintive, valutando le sue prestazioni su diversi dataset e confrontandolo con altri modelli di segmentazione. L'obiettivo è comprendere a fondo i meccanismi che stanno alla base del successo di SAM e individuare potenziali direzioni future per lo sviluppo di modelli di segmentazione sempre più sofisticati.

Capitolo 1. Cos'è SAM?

1.1 Introduzione



Panoramica di SAM [1].

Immaginiamo di avere un'immagine e di voler selezionare una parte specifica, come un oggetto o una persona. Il modello SAM è progettato proprio per fare questo in modo automatico e preciso.

SAM guardando un'immagine riesce a identificare e isolare i diversi elementi. In termini più tecnici, è composto da tre parti principali:

1. Image encoder: Questa parte "studia" l'immagine, creando una rappresentazione interna che cattura le caratteristiche più importanti.
2. Prompt encoder: Qui entrano in gioco le nostre richieste. Si può indicare un punto preciso, un rettangolo o una descrizione, e il modello trasformerà queste informazioni in un formato che può comprendere.
3. Mask decoder: Questa parte finale crea la maschera di segmentazione, ovvero una sorta di "stampo" che indica esattamente quali pixel dell'immagine appartengono all'oggetto che abbiamo selezionato.

SAM è un modello all'avanguardia che rappresenta un grande passo avanti nel campo della visione artificiale. La sua capacità di segmentare immagini in modo preciso e versatile lo rende uno strumento prezioso in molti ambiti, dall'editing fotografico alla medicina.

1.1.1 Image Encoder

Il suo compito è analizzare l'intera immagine e creare una rappresentazione numerica, o "embedding", che cattura le caratteristiche più importanti dell'immagine stessa. È come se stesse creando una sorta di "impronta digitale" dell'immagine.

Come funziona:

- Vision Transformer (ViT): Utilizza una rete neurale chiamata Vision Transformer, che è particolarmente brava a "capire" le relazioni tra diverse parti dell'immagine. Immagina di dividere l'immagine in tanti piccoli pezzi (patch). ViT analizza queste patch e le mette in relazione tra loro, creando una rappresentazione più astratta e informativa dell'intera immagine.
- Alta risoluzione: SAM è progettato per funzionare con immagini ad alta risoluzione. Per questo motivo, il ViT utilizzato è stato adattato per gestire input di grandi dimensioni.
- Output: L'output del codificatore è un tensore (array multidimensionale), dove ogni dimensione rappresenta una caratteristica dell'immagine. Questo tensore viene poi utilizzato dal decoder per generare la maschera di segmentazione.

1.1.2 Prompt Encoder

Il suo compito è trasformare le nostre richieste (punti, box, descrizioni) in un formato numerico che il modello possa comprendere.

Come funziona:

- Rappresentazione vettoriale: Ogni tipo di prompt viene rappresentato da un vettore numerico. Ad esempio, un punto viene rappresentato da un vettore che indica le sue coordinate nell'immagine.
- Codifica posizionale: Per i prompt spaziali (punti, box), viene aggiunta una codifica posizionale che indica la posizione del prompt all'interno dell'immagine.
- Codifica testuale: Per i prompt testuali, viene utilizzato un modello pre-addestrato (come CLIP) per ottenere una rappresentazione numerica del testo.

CLIP è un modello di intelligenza artificiale che ha imparato a collegare le parole con le immagini. Ad esempio, se gli viene mostrata un'immagine di un cane, CLIP sarà in grado di associarla alle parole "cane", "animale", "cucciolo" e così via.

1.1.3 Mask Decoder

Utilizza l’embedding dell'immagine e i prompt codificati per generare la maschera di segmentazione.

Come funziona:

- Attenzione: Il decoder utilizza un meccanismo di attenzione per focalizzarsi sulle parti dell'immagine più rilevanti per la tua richiesta. Questo significa che il decoder "guarda" attentamente le parti dell'immagine che corrispondono al prompt.
- Iterazioni: Il processo di decodifica avviene in più passaggi, ad ogni passaggio il decoder rende la maschera di segmentazione più precisa e dettagliata.
- Output: Alla fine, il decoder produce una maschera binaria, dove i pixel con valore 1 appartengono all'oggetto da segmentare e quelli con valore 0 non appartengono all'oggetto.

1.1.4 Risolvere le ambiguità

Quando si fornisce al modello SAM un prompt ambiguo, cioè una richiesta che potrebbe corrispondere a più oggetti nell'immagine, il modello potrebbe generare più maschere valide. Per affrontare questo problema, è stato modificato il modello affinché preveda più maschere di output per un singolo prompt.

Per aiutare l'utente a scegliere la maschera più corretta, SAM assegna a ciascuna di esse un punteggio di confidenza**.** Questo punteggio indica quanto il modello è sicuro che la maschera corrisponda effettivamente all'oggetto cercato e rappresenta la stima dell'IoU tra la maschera prevista e la maschera ground truth.

1.1.5 Efficienza

La progettazione complessiva del modello è fortemente orientata all'efficienza. Dato un embedding pre-calcolato dell'immagine, il prompt encoder e il mask decoder vengono eseguiti in un browser Web, su CPU, in 50 millisecondi. Queste prestazioni di runtime consentono un'interazione fluida e in tempo reale con SAM.

1.2 Data engine

Uno dei maggiori ostacoli nello sviluppo di un modello di segmentazione come SAM è la carenza di dati di alta qualità. Le maschere di segmentazione, non sono così abbondanti come altri tipi di dati visivi. Per risolvere questo problema, è stato creato un motore di dati appositamente progettato per raccogliere un enorme dataset di maschere (1,1 miliardi), chiamato SA-1B.

Il data engine ha tre fasi: (1) assisted-manual stage, (2) semi-automatic stage e (3) fully-automatic stage.

1.2.1 Assisted-manual stage

In questa fase iniziale, un team di annotatori esperti creava manualmente le maschere, cliccando sui pixel che appartenevano all'oggetto da segmentare. Per rendere il processo più efficiente, hanno utilizzato uno strumento interattivo basato su browser che sfruttava le capacità di SAM per fornire suggerimenti in tempo reale. Le maschere venivano perfezionate utilizzando strumenti “brush” ed “eraser” precisi al pixel. Non sono stati imposti vincoli semantici per l’etichettatura degli oggetti e gli annotatori hanno etichettato liberamente sia “stuff” che “things”. Agli annotatori è stato chiesto di etichettare gli oggetti in ordine di importanza e sono stati incoraggiati a procedere all’immagine successiva una volta che la maschera impiegava più di 30 secondi per annotarla.

All’inizio di questo stage, SAM è stato addestrato utilizzando dataset di segmentazione pubblici comuni. Dopo una sufficiente annotazione dei dati, SAM è stato riaddestrato utilizzando solo le maschere appena annotate. Man mano che venivano raccolte più maschere, l’architettura dell’image encoder veniva aggiornata da ViT-B a ViT-H; in totale il modello è stato riaddestrato 6 volte. Il tempo medio di annotazione per maschera è diminuito da 34 a 14 secondi e il numero medio di maschere per immagine è aumentato da 20 a 44. Nel complesso, in questa fase sono state raccolte 4,3 milioni di maschere da 120.000 immagini.

1.2.2 Semi-automatic stage

In questa fase, l'obiettivo era aumentare la diversità delle maschere per migliorare la capacità del modello di segmentare qualsiasi oggetto. Per concentrare gli annotatori sugli oggetti meno evidenti, è stato inizialmente addestrato un rilevatore di oggetti generico su tutte le maschere del primo stage, utilizzando una categoria generica di "object". Successivamente, le immagini sono state precompilate con le maschere individuate automaticamente dal rilevatore e presentate agli annotatori, che dovevano etichettare gli oggetti mancanti.

Durante questa fase sono state raccolte ulteriori 5,9 milioni di maschere in 180.000 immagini, portando il totale a 10,2 milioni. Come nel primo stage, il modello è stato periodicamente riaddestrato sui nuovi dati (5 volte). Il tempo medio di annotazione per maschera è tornato a 34 secondi, principalmente a causa della complessità degli oggetti rimanenti. Il numero medio di maschere per immagine è aumentato da 44 a 72, grazie anche al contributo delle maschere generate automaticamente.

1.2.3 Fully automatic stage

Nella fase finale, l'annotazione è diventata completamente automatica. Ciò è stato possibile grazie a due importanti miglioramenti apportati al modello. Innanzitutto, all'inizio di questa fase, era stato raccolto un dataset sufficientemente ampio per addestrare un modello molto preciso. In secondo luogo, era stato sviluppato un modello in grado di gestire situazioni ambigue, in cui un pixel potrebbe appartenere a più oggetti. Con il modello sensibile all'ambiguità, se un punto si trova su una parte o sotto parte, il modello restituirà la sotto parte, la parte e l'intero oggetto. Per esempio, se il soggetto è una persona che tiene in mano una tazza, la parte può essere la tazza e la sotto parte può essere il manico. Verrebbe quindi restituita la tazza, il manico e la persona che tiene in mano la tazza.

Nello specifico, il modello generava delle proposte di maschere a partire da una griglia regolare sovrapposta all'immagine. Per ciascuna posizione della griglia, il modello prevedeva un insieme di possibili maschere. Successivamente, il modello selezionava le maschere più probabili e precise utilizzando la metrica IoU e verificando la loro stabilità rispetto a piccole variazioni della soglia di probabilità. Infine, venivano eliminate le maschere ridondanti.

Per migliorare la qualità delle maschere, soprattutto per gli oggetti più piccoli, sono state elaborate anche delle versioni ingrandite delle immagini. Questo processo è stato applicato a tutte le 11 milioni di immagini del dataset, generando un totale di 1,1 miliardi di maschere di alta qualità.

1.3 Dataset

Il dataset, SA-1B, è costituito da 11 milioni di immagini diverse, ad alta risoluzione, con licenza e che proteggono la privacy e da 1,1 miliardi di maschere di segmentazione di alta qualità raccolte con il data engine. SA-1B è stato rilasciato per favorire lo sviluppo futuro di modelli di base per la visione artificiale. Di seguito verrà confrontato con i dataset esistenti e verranno analizzati la qualità e le proprietà delle maschere.

1.3.1 Immagini

Gli autori hanno concesso in licenza un nuovo set di 11 milioni di immagini da un fornitore che lavora direttamente con i fotografi. Queste immagini sono ad alta risoluzione (3300 x 4950 pixel in media) e la dimensione dei dati risultante può presentare sfide in termini di accessibilità e archiviazione. Pertanto, sono state rilasciate immagini sottocampionate con il lato più corto impostato su 1550 pixel. Anche dopo il sottocampionamento, le immagini hanno una risoluzione significativamente più elevata rispetto a molti dataset sulla visione artificiale esistenti (ad esempio, le immagini del dataset

COCO sono 480 x 640 pixel). Bisogna inoltre tener presente che la maggior parte dei modelli oggi funziona con ingressi a risoluzione molto più bassa. Nelle immagini pubblicate i volti e le targhe dei veicoli sono stati sfocati.

1.3.2 Qualità della maschera

Per stimare la qualità della maschera, sono state campionate casualmente 500 immagini (circa 50.000 maschere) ed è stato chiesto agli annotatori professionisti di migliorare la qualità di tutte le maschere di queste immagini. Gli annotatori lo hanno fatto utilizzando il modello e gli strumenti di modifica "brush" ed "eraser" precisi al pixel. Questa procedura ha prodotto coppie di maschere previste automaticamente e corrette professionalmente. Hanno calcolato IoU tra ciascuna coppia ed è stato scoperto che il 94% delle coppie ha più del 90% IoU (e il 97% delle coppie ha più del 75% IoU). Per fare un confronto, il lavoro precedente [7, 8] stimava la coerenza tra annotatori all'85%-91% IoU.

1.3.3 Proprietà delle maschere

Gli autori hanno analizzato la distribuzione spaziale degli oggetti nei diversi dataset e hanno scoperto che SA-1B presenta una maggiore copertura degli angoli dell'immagine rispetto a LVIS v1 [7] e ADE20K [9]. Ciò significa che gli oggetti in SA-1B sono distribuiti in modo più uniforme su tutta l'area dell'immagine, inclusi i bordi, mentre in LVIS v1 e ADE20K si concentra una maggiore densità di oggetti nella parte centrale dell'immagine. Questo fenomeno è noto come bias centrale.

Al contrario, COCO e Open Images V5 mostrano un bias centrale più prominente, con una concentrazione di oggetti nella parte centrale delle immagini.

Per quanto riguarda le dimensioni, SA-1B supera di gran lunga tutti gli altri dataset considerati. Con oltre 11 milioni di immagini e 1,1 miliardi di maschere, SA-1B ha 11 volte il numero di immagini di Open Images V5 e 400 volte il numero di maschere di Open Images V5, il secondo dataset più grande. In media, ogni immagine in SA-1B contiene 36 volte più maschere rispetto a Open Images V5. Anche rispetto ad ADE20K, il dataset più simile per quanto riguarda la distribuzione delle maschere, SA-1B presenta ancora un numero di maschere 3,5 volte superiore.

1.4 Analisi RAI

Gli autori hanno eseguito un'analisi RAI (Responsible AI) del loro lavoro indagando su potenziali problemi di equità e pregiudizi quando si utilizzano SA-1B e SAM. Si sono concentrati sulla distribuzione geografica e del reddito di SA-1B e sull'equità di SAM rispetto agli attributi protetti delle persone (genere, età, etnia, origine nazionale).

SA-1B ha una percentuale di immagini sostanzialmente più elevata in Europa, Asia e Oceania, nonché nei paesi a reddito medio. Tutti i dataset sottorappresentano l'Africa e i paesi a basso reddito. Si nota che in SA-1B, tutti i continenti, compresa l'Africa, hanno almeno 28 milioni di maschere, 10 volte in più rispetto al numero totale di maschere di qualsiasi dataset precedente. Infine, si nota che il numero medio di maschere per immagine è abbastanza coerente tra regioni e redditi (94 – 108 per immagine).

Hanno usato il dataset More Inclusive Annotations for People (MIAP [10]) per la presentazione del genere e l'età e un dataset proprietario per il tono della pelle. È stato notato che le donne sono sottorappresentate nei dataset di rilevamento e segmentazione, ma SAM si comporta in modo simile tra i gruppi. Hanno ripetuto l'analisi per l'età percepita, notando che coloro che sono percepiti come più giovani e più anziani hanno dimostrato di essere sottorappresentati in dataset su larga scala. SAM offre risultati migliori su coloro che sono percepiti come più anziani (sebbene l'intervallo di confidenza sia più ampio). Infine, hanno ripetuto l'analisi per la tonalità della pelle percepita, notando che quelli con tonalità della pelle più chiara sono sovrarappresentati e quelli con tonalità della pelle più scura sottorappresentati in dataset su larga scala.

1.5 Esperimenti di trasferimento zero-shot

Gli autori considerano cinque attività, quattro delle quali differiscono significativamente dall'attività di segmentazione da prompt utilizzata per addestrare SAM. Questi esperimenti valutano il modello su dataset e attività che non sono stati visti durante l'addestramento. Per zero-shot si intende la capacità di un modello di machine learning di eseguire un compito per il quale non è stato specificamente addestrato. In questo caso, SAM, pur essendo stato addestrato su un'attività specifica (segmentazione da prompt), viene valutato su altre attività completamente nuove, senza aver visto alcun esempio durante la fase di addestramento. Questo dimostra la generalizzabilità del modello e la sua capacità di adattarsi a nuovi scenari. Il transfer learning, invece, si riferisce alla capacità di un modello di applicare conoscenze acquisite su un compito a un compito diverso, ma correlato.

I loro esperimenti esplorano sia lo scenario dello zero-shot che quello del transfer learning.

1.5.1 Valutazione della maschera valida a punto singolo in zero-shot

Compito: Valutare la segmentazione di un oggetto da un unico punto in primo piano. Questo compito è mal posto poiché un punto può fare riferimento a più oggetti. Le maschere di verità nella maggior parte dei dataset non enumerano tutte le maschere possibili, il che può rendere inaffidabili le metriche automatiche. Pertanto, viene integrata la metrica mIoU standard (ovvero la media su tutti gli IoU tra le maschere previste e quelle di base) con uno studio umano in cui gli annotatori valutano la qualità della maschera da 1 (senza senso) a 10 (perfetto al pixel).

Utilizzano una nuova lista di 23 dataset (presi casualmente tra dataset esistenti) per la valutazione mIoU.

Per lo studio umano, viene utilizzato un sottoinsieme di questi dataset (a causa dei requisiti di risorse di tale studio). Questo sottoinsieme include entrambi i dataset per i quali SAM ha prestazioni superiori e inferiori a RITM secondo i parametri automatici.

Innanzitutto si esamina la valutazione automatica sull'intera suite di 23 dataset utilizzando mIoU. SAM produce risultati più elevati su 16 dei 23 dataset, fino a circa 47 IoU.

Riferendosi allo studio umano, le valutazioni medie di SAM sono comprese tra 7 e 9, che corrisponde alle linee guida di valutazione qualitativa: "*Un punteggio alto (7-9): l'oggetto è identificabile e gli errori sono piccoli e rari (ad esempio, manca un piccolo componente disconnesso fortemente oscurato, ...).*". Questi risultati indicano che SAM ha imparato a segmentare le maschere valide da un singolo punto.

1.5.2 Rilevamento del bordo a zero-shot

Viene valutato SAM sul classico compito di basso livello di rilevamento dei bordi utilizzando BSDS500 [11, 12]. Viene utilizzata una versione semplificata della loro pipeline di generazione automatica delle maschere. SAM viene stimolato con una griglia di 16x16 punti per generare 768 maschere iniziali. Le maschere ridondanti vengono rimosse.

Quantitativamente si osserva che, anche se SAM non è stato addestrato per il rilevamento dei bordi, produce mappe dei bordi ragionevoli. Rispetto alla ground truth, SAM prevede più bordi, compresi quelli sensibili, ovvero contorni nei quali una piccola variazione nella loro posizione o forma può influenzare significativamente la qualità della segmentazione di un'immagine, che non sono annotati in BSDS500. Inoltre, SAM funziona bene rispetto ai metodi pionieristici di deep learning come HED (anche addestrati su BSDS500) e significativamente migliore rispetto ai metodi di trasferimento zero-shot precedenti.

1.5.3 Proposte di oggetti zero-shot

Per valutare la capacità di SAM di generare proposte di oggetti, gli autori hanno adattato leggermente la loro pipeline di generazione automatica delle maschere. Le proposte di oggetti vengono calcolate applicando una soglia alle maschere generate da SAM. In pratica, si considera come proposta ogni regione connessa di pixel che supera una certa soglia di confidenza.

ViTDeT [13] è un modello di rilevamento di oggetti basato sull'architettura ViT (Vision Transformer), progettato per identificare e localizzare oggetti all'interno di immagini. La sua struttura semplice e lineare lo rende particolarmente efficace nell'estrazione di caratteristiche visive di alto livello.

Per misurare le prestazioni dei diversi modelli, è stata utilizzata la metrica standard di average recall (AR) sul dataset LVIS v1. LVIS è stato scelto per la sua complessità, data l'elevata varietà di categorie di oggetti presenti.

Come ci si aspettava, l'utilizzo delle proposte generate da ViTDet-H ha portato ai migliori risultati complessivi. Tuttavia, SAM si è dimostrato particolarmente competitivo in diverse situazioni. In particolare, ha superato ViTDet-H nel rilevamento di oggetti di medie e grandi dimensioni, nonché di oggetti sia rari che comuni. Solo nel caso di oggetti piccoli e molto frequenti, ViTDet-H ha ottenuto risultati migliori, probabilmente a causa di un adattamento più specifico alle peculiarità del dataset LVIS, su cui è stato addestrato. SAM, non essendo stato addestrato su LVIS, risulta meno influenzato da eventuali bias presenti nelle annotazioni.

1.5.4 Zero-shot text-to-mask

Infine, si considera un compito di livello elevato: segmentare oggetti da testo in formato libero. Questo esperimento è una prova della capacità di SAM di elaborare istruzioni di testo. Anche se hanno utilizzato esattamente lo stesso SAM in tutti gli esperimenti precedenti, per questo la procedura di addestramento di SAM è stata modificata per renderlo sensibile al testo, ma in un modo che non richiede annotazioni di testo.

Risultati: SAM può segmentare gli oggetti in base a semplici prompt di testo come “una ruota” e a frasi come “griglia a dente di castoro”. Quando SAM non riesce a scegliere l’oggetto giusto solo da un messaggio di testo, un punto aggiuntivo spesso corregge la previsione.

Capitolo 2. Come utilizzare SAM?

2.1 Riallenare parzialmente la rete, SAMUS

2.1.1 Introduzione a SAMUS

SAMUS [5] è un modello universale su misura per la segmentazione delle immagini ecografiche. Rispetto ai precedenti modelli di base basati su SAM, SAMUS si concentra maggiormente sull'integrazione delle funzionalità locali e sulla riduzione del consumo di GPU, che è fondamentale per una segmentazione delle immagini mediche accurata e facile da implementare negli scenari clinici.

I modelli tradizionali per la segmentazione delle immagini mediche sono spesso specializzati per compiti specifici e hanno difficoltà a generalizzare a nuovi dati. Inoltre, richiedono spesso un intervento manuale per definire le regioni di interesse.

Questo nuovo modello risolve questi problemi introducendo:

- Un ramo parallelo: Aggiunge dettagli locali all'immagine, migliorando la precisione della segmentazione.
- Adattatori: Consentono al modello di adattarsi alle caratteristiche specifiche delle immagini ecografiche.
- Generatore di prompt automatico: Elimina la necessità di un intervento manuale, rendendo il processo completamente automatico.

Per la verifica viene raccolto un dataset ecografico completo, comprendente circa 30000 immagini e 69000 maschere e che copre sei categorie di oggetti.

2.1.2 Visual tuning

Negli ultimi anni, l'avvento dei modelli di base nella visione artificiale ha rivoluzionato il modo in cui vengono affrontati i problemi di analisi delle immagini. Questi modelli, pre-addestrati su enormi dataset, offrono una base solida per lo sviluppo di applicazioni specifiche.

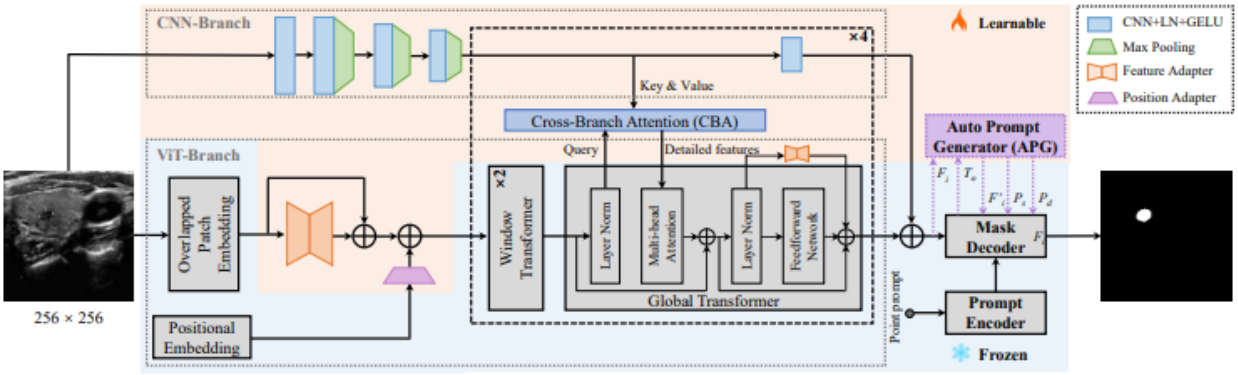
Per adattare questi modelli pre-addestrati a compiti particolari, sono state proposte diverse strategie di "ottimizzazione visiva".

Le tecniche di ottimizzazione visiva si possono suddividere in cinque categorie principali:

- Fine-tuning: In questo caso, viene "specializzato" l'intero modello o solo alcune sue parti per adattarlo al nuovo compito tramite la regolazione dell'intero set di parametri dei modelli pre-addestrati.
- Parameter tuning: Qui, si interviene direttamente sui parametri interni del modello, modificando i pesi o i bias.
- Remapping tuning: In questa tecnica, vengono trasferite le conoscenze acquisite da un modello pre-addestrato a un altro modello.
- Prompt tuning: Questa strategia è più innovativa e consiste nell'aggiungere al modello delle informazioni specifiche sul compito da svolgere, come delle istruzioni dettagliate.
- Adapter tuning: Questa è una delle tecniche più popolari e consiste nell'aggiungere dei "moduli aggiuntivi" al modello pre-addestrato, che gli permettono di apprendere rapidamente nuovi compiti.

Ognuna di queste tecniche presenta vantaggi e svantaggi, e la scelta della strategia più adatta dipende dal compito specifico, dalle risorse disponibili e dalle caratteristiche del modello pre-addestrato.

2.1.3 Architettura



Panoramica di SAMUS [5].

Come già accennato, SAMUS eredita la struttura fondamentale di SAM. Tuttavia, per adattarlo al mondo specifico delle immagini mediche, sono state introdotte una serie di modifiche mirate all'image encoder:

- Riduzione della risoluzione**:** La dimensione delle immagini in ingresso è stata ridotta, passando da 1024x1024 pixel a 256x256 pixel. Questa scelta permette di ridurre significativamente il carico computazionale e la memoria necessaria per processare le immagini, rendendo il modello più efficiente.

- Sovrapposizione dell'embedding della patch: La sovrapposizione delle patch significa che lo stesso pixel sarà presente in più vettori, contribuendo così a una rappresentazione più ricca e informativa dell'immagine. Questa tecnica permette di estrarre informazioni più dettagliate dalle immagini, migliorando la capacità del modello di identificare le piccole strutture presenti nelle immagini mediche.
- Adattatori al ramo ViT: Sono stati introdotti degli adattatori specifici per il ramo ViT. Questi adattatori permettono al modello di adattarsi meglio alle caratteristiche delle immagini mediche.
- Ramo CNN parallelo e attenzione al ramo incrociato (CBA): Hanno aggiunto un ramo parallelo basato su una rete neurale convoluzionale (CNN) per estrarre informazioni locali dalle immagini. Queste informazioni vengono poi integrate con quelle globali estratte dal ramo ViT grazie al meccanismo di attenzione al ramo incrociato. In pratica, il modello impara a "prestare attenzione" sia alle caratteristiche globali dell'immagine (fornite dal ramo ViT) sia a quelle locali (fornite dal ramo CNN), ottenendo così una rappresentazione più completa e accurata dell'immagine.

Il risultato finale è un modello più efficiente e accurato per la segmentazione delle immagini ecografiche.

2.1.4 Training

Prima dell'addestramento, SAMUS inizializza i parametri ereditati da SAM utilizzando i pesi addestrati su SA-1B. I restanti parametri vengono inizializzati in modo casuale. Durante il processo di training, vengono aggiornati solo i parametri degli adattatori, del ramo CNN e del modulo CBA, mentre gli altri parametri vengono mantenuti congelati.

2.1.5 Esperimenti e comparazione

Per valutare in modo completo l'efficacia di SAMUS, è stato costruito un ampio dataset denominato US30K, contenente dati provenienti da sette dataset disponibili al pubblico, inclusi TN3K, DDTI, TG3K, BUSI, UDIAT, CAMUS e HMCQU. Questo dataset è composto da 30106 immagini e 68570 maschere, dati suddivisi in 6 categorie: nodulo tiroideo, ghiandola tiroidea, tumore al seno, ventricolo sinistro, miocardio e atrio sinistro. I dati di TN3K e TG3K sono suddivisi in set di training, validazione e test, BUSI è suddiviso casualmente in 7:1:2 rispettivamente per training, validazione e test, CAMUS è suddiviso prima in un set di training e in un set di prova a seconda della sfida. Quindi, vengono selezionati casualmente il 10% dei pazienti dal set di training per convalidare il modello e i dati rimanenti sono utilizzati come dati di allenamento finali. Per valutare la generalizzazione dei

diversi modelli, gli altri dataset in US30K non vengono visualizzati durante il processo di formazione e convalida.

Comparando SAMUS con dodici approcci SOTA specifici sui database TN3K, BUSI e CAMUS (CAMUS-LV, CAMUS-MYO e CAMUS-LA), si nota che SAMUS ottiene Dice score e HD (Hausdorff Distance) migliori su TN3K, BUSI e CAMUS-LA, mentre su CAMUS-LV e CAMUS-MYO ottiene Dice score migliori e HD inferiori.

SAMUS viene comparato, poi, con quattro foundation models SOTA: SAM, MedSAM, SAMed e MSA. Il modello addestrato su SA-1B mostra un significativo degrado delle prestazioni sulla segmentazione delle immagini mediche (nei cinque dataset, TN3K, BUSI, CAMUS-LV, CAMUS-MYO e CAMUS-LA, ottiene i seguenti Dice score, rispettivamente, di 29.59%, 54.01%, 28.18%, 29.42% e 17.28%). Con una semplice fine-tuning del mask decoder di SAM basato sul dataset US30K, MedSAM migliora notevolmente le prestazioni di SAM (nei cinque dataset ottiene i seguenti Dice score, rispettivamente, di 71.09%, 77.75%, 87.52%, 76.07% e 88.06%). MAS, il modello con le migliori prestazioni trai i foundation models di confronto, migliora efficacemente le prestazioni di segmentazione di SAM con un aumento in media di 53.08%, 27.65%, 62.77%, 53.05%, e 74.52% nel Dice score. Rispetto a MSA, SAMUS ottiene costantemente notevoli miglioramenti nei cinque dataset, con Dice score superiori rispettivamente di 83.05%, 84.54%, 91.13%, 83.11% e 92%. Questo convalida l'efficacia del ramo CNN e del modulo CBA in SAMUS, soprattutto per integrare le informazioni locali che sono cruciali per la segmentazione delle immagini mediche.

Inoltre, confrontando SAMUS con SAM dal punto di vista del costo della memoria GPU, si vede che quello di SAMUS è solo il 28% di quello richiesto per addestrare l'intero SAM.

2.2 Prompt engineering

2.2.1 Introduzione al prompt engineering

La visual prompt engineering consiste nella progettazione e ottimizzazione di sequenze di token, definite come prompt, per guidare modelli di linguaggio di grandi dimensioni verso la generazione di output desiderati. In ambito NLP, questa tecnica è impiegata per adattare in modo efficiente modelli pre-addestrati a compiti specifici, sfruttando il meccanismo di fine-tuning dei parametri.

I prompt possono essere classificati in base alla loro posizione all'interno del testo:

- Prompt cloze: Inseriti all'interno del testo.
- Prompt prefix: Aggiunti alla fine del testo.

La creazione manuale di prompt, sebbene intuitiva, presenta limitazioni in termini di efficienza e scalabilità. Per superare queste limitazioni, sono stati proposti approcci di apprendimento automatico dei prompt:

- Prompt discreti: Vengono cercati nello spazio discreto dei token, tipicamente attraverso algoritmi di ricerca basati su gradiente.
- Prompt continui: Vengono rappresentati come vettori continui nello spazio degli embedding, consentendo un'ottimizzazione più fine dei parametri.

I vantaggi della prompt engineering sono:

- Flessibilità: Adattamento efficiente di modelli pre-addestrati a nuovi compiti.
- Efficienza: Riduzione del numero di parametri da addestrare rispetto al fine-tuning completo.
- Miglioramento delle prestazioni: Ottenimento di risultati competitivi su una vasta gamma di compiti.

In sintesi, l'ingegneria dei prompt rappresenta una promettente direzione di ricerca nell'ambito del NLP, consentendo di sfruttare al meglio le potenzialità dei modelli di linguaggio di grandi dimensioni.

2.2.2 Rilevamento degli oggetti

Nonostante SAM affermi la sua capacità di segmentare qualsiasi oggetto, la sua applicazione pratica è stata messa in discussione. In particolare, sono state sollevate preoccupazioni riguardo all'efficacia di SAM [19] per applicazioni quali la segmentazione di immagini mediche [2, 14, 17, 18, 20, 21], il rilevamento di oggetti mimetici, il rilevamento di oggetti trasparenti e specchiati e altri scenari simili. Di conseguenza, studi recenti si sono concentrati sulla valutazione delle prestazioni del SAM in vari contesti. Questi studi hanno dimostrato che i prompt di punti o di box sono altamente efficaci in vari scenari pratici. SAM ha ottenuto solide prestazioni zero-shot nei settori delle immagini naturali, del telerilevamento e dell'imaging medico [14, 15, 16, 17, 18]. Tuttavia, la sua capacità di generalizzare in scenari applicativi complessi, in particolare dove le informazioni semantiche sono ambigue o in ambienti con basso contrasto, potrebbe non soddisfare i requisiti delle attività. Pertanto, sono necessarie ulteriori ricerche per migliorare le prestazioni di SAM in ambienti complessi. Nel campo del rilevamento dei crateri, SAM viene sfruttato per la segmentazione automatizzata delle immagini. Successivamente, viene valutata la forma di ciascuna maschera segmentata e vengono eseguite anche ulteriori fasi di elaborazione, come il filtraggio e l'estrazione dei confini.

Integrando la prompt engineering con l'architettura modulare autoadattativa di SAM, è possibile selezionare i metodi di rilevamento degli oggetti più adatti in base a compiti e ambienti diversi, ottenendo così capacità di rilevamento degli oggetti più intelligenti e flessibili.

2.2.3 Conteggio degli oggetti

Nel campo del conteggio degli oggetti, i ricercatori [22, 23] adottano SAM impiegando box di delimitazione come prompt per generare maschere di segmentazione. Le caratteristiche dell'immagine densa, una rappresentazione dell'immagine che cattura un'informazione molto ricca e dettagliata su ogni pixel o regione dell'immagine, ottenuta dall'image encoder vengono moltiplicate e viene calcolata la media con il vettore delle caratteristiche di un oggetto di riferimento. Successivamente, come prompt per la segmentazione viene utilizzata una griglia di punti composta da 32 punti per bordo. Il vettore delle caratteristiche della maschera risultante si ottiene moltiplicandolo e facendo la media insieme alle caratteristiche dense, ovvero rappresentazioni numeriche di un'immagine che cattura un'enorme quantità di dettagli su ogni pixel o regione dell'immagine (colore, texture, forma). Alla fine, per determinare il conteggio totale, viene calcolata la somiglianza del coseno (una misura della somiglianza tra due vettori definita in termini del coseno dell'angolo tra di loro) tra la maschera prevista e i vettori delle caratteristiche dell'esempio di riferimento. Quando la somiglianza del coseno supera una soglia predefinita, l'oggetto target viene considerato riconosciuto. Calcolando tutti gli oggetti target si ottiene infine il conteggio totale.

2.2.4 Telerilevamento

Nell'ambito della segmentazione delle immagini telerilevate [24, 25], a causa della prospettiva dall'alto verso il basso delle immagini telerilevate, gli oggetti all'interno della scena possono avere diverse orientazioni, cioè possono essere ruotati in modi diversi. Di conseguenza, è stata proposta una tecnica per utilizzare il rettangolo orizzontale che racchiude il minimo box di delimitazione ruotato (R-Box) come guida per la segmentazione di SAM durante la progettazione dei prompt. Per il prompt della maschera, viene definita come l'area corrispondente racchiusa dalla box di delimitazione [26].

2.2.5 SAM adapter

SAM-Adapter [27] è stato sviluppato per arricchire il modello SAM originale con conoscenze specializzate in un determinato dominio, ad esempio quello medico, migliorandone significativamente la capacità di generalizzare a diverse attività. Questa capacità di generalizzare è cruciale per rendere SAM più versatile e adattabile a una vasta gamma di applicazioni. L'adattatore funziona acquisendo conoscenze rilevanti e generando prompt specifici per ogni attività durante la fase iniziale. Grazie a questi prompt più informativi, la rete segmentata di SAM ottiene prestazioni

notevolmente superiori in compiti complessi come il rilevamento di oggetti pseudocolori, il rilevamento di ombre e la segmentazione di immagini mediche.

2.2.6 Text2Seg

Text2Seg [25] introduce un modello di linguaggio visivo che si basa su istruzioni di testo come input. Il modello funziona come segue: innanzitutto, il prompt di testo funge da input, che genera riquadri di delimitazione, questi riquadri di delimitazione guidano SAM nella generazione delle maschere di segmentazione, successivamente, il processo CLIP Surgery genera mappe di calore, rappresentazioni visive che utilizzano un gradiente di colore per mostrare la distribuzione di una particolare caratteristica o valore in un'immagine (in questo caso indicano la probabilità che un pixel appartenga a una particolare classe o categoria), utilizzando i prompt di testo e i prompt dei punti derivati da queste mappe di calore vengono inseriti in SAM e infine, viene applicato un algoritmo di similarità per ottenere la mappa di segmentazione definitiva.

2.2.7 SAMText

SAMText [28] introduce una metodologia versatile per generare maschere di segmentazione mirate al testo della scena in immagini o fotogrammi video. Il processo inizia, una volta fornito l'input, estraendo le coordinate della box di delimitazione da un modello di rilevamento del testo della scena, utilizzando le annotazioni esistenti. Queste coordinate della box di delimitazione estratte fungono da prompt per SAM, che facilita la successiva generazione di maschere. Se i riquadri di delimitazione mostrano un orientamento, SAMText calcola i loro box di delimitazione minimi per ottenere box di delimitazione orizzontali, che, a loro volta, servono come prompt di SAM per la generazione della maschera.

2.2.8 Applicazione su più domini

Visual prompt e modelli visivi di grandi dimensioni hanno consentito progressi significativi in campi in cui la comprensione e l'analisi visiva sono fondamentali. Ad esempio, SAM basato su prompt ha sbloccato nuove opportunità in settori quali l'imaging medico, l'agricoltura, l'editing di immagini, il rilevamento di oggetti, la localizzazione audiovisiva e altro ancora [22]. Nel settore medico, visual prompts come maschere di segmentazione, box di delimitazione e punti chiave vengono utilizzati per aiutare a rilevare malattie, quantificare la gravità delle lesioni e analizzare le scansioni mediche [2, 21, 29, 30]. Per le tipiche procedure di trattamento in radioterapia oncologica, hanno confrontato i risultati di Dice e Jaccard tra la delineazione manuale clinica e la segmentazione automatica utilizzando SAM con box prompt e hanno dimostrato le solide capacità di generalizzazione di SAM nella segmentazione automatica per la radioterapia [29]. In agricoltura, i visual prompt potrebbero

essere utilizzati per monitorare la crescita dei raccolti, rilevare erbe infestanti o parassiti e stimare i raccolti [22, 31]. In un esperimento sono state valutate le prestazioni di segmentazione zero-shot di SAM su compiti rappresentativi di segmentazione dei polli ed è stato dimostrato che il tracciamento degli oggetti basato su SAM potrebbe fornire dati preziosi sul comportamento e sui modelli di movimento dei polli [32].

2.3 SAM per l'analisi delle immagini mediche: 1° esperimento

2.3.1 Introduzione

La segmentazione delle immagini è un compito centrale nell'analisi delle immagini mediche , che va dalla segmentazione di organi, anomalie, ossa e altri e che ha ricevuto progressi significativi dal deep learning. Tuttavia, lo sviluppo e l'addestramento di modelli di segmentazione per nuovi dati e/o attività di imaging medico è praticamente impegnativo, a causa della natura costosa e dispendiosa in termini di tempo della raccolta e della cura delle immagini mediche, principalmente perché i radiologi addestrati devono in genere fornire attente annotazioni sulle maschere per le immagini.

Queste difficoltà potrebbero essere significativamente mitigate con l'avvento dei foundation models e dell'apprendimento zero-shot. SAM, come precedentemente detto, ha raggiunto promettenti prestazioni di segmentazione zero-shot su una varietà di dataset di immagini naturali.

2.3.2 Come segmentare le immagini mediche con SAM?

SAM è progettato per richiedere un prompt o una serie di prompt per produrre una maschera di segmentazione. Tecnicamente, può essere eseguito senza la richiesta di fornire alcun oggetto visibile, ma ciò è inutile per le immagini mediche, dove spesso sono presenti molti altri oggetti nell'immagine oltre a quello di interesse. Data questa natura basata su prompt, nella sua forma base, SAM non può essere utilizzato allo stesso modo della maggior parte dei modelli di segmentazione nell'imaging medico in cui l'input è semplicemente un'immagine e l'output è una maschera di segmentazione o più maschere per l'oggetto o gli oggetti desiderati.

Vengono proposti tre modi principali in cui SAM può essere utilizzata nel processo di segmentazione delle immagini mediche. I primi due prevedono l'utilizzo dell'effettivo Segment Anything Model nel processo di annotazione, generazione di maschere o addestramento di modelli aggiuntivi. Questi approcci non comportano modifiche a SAM. Il terzo approccio prevede il processo di formazione/ottimizzazione di un modello simile a SAM mirato alle immagini mediche.

Annotazione semiautomatica ("human in the loop"). L'annotazione manuale delle immagini mediche è una delle principali sfide nello sviluppo di modelli di segmentazione in questo campo poiché

richiede in genere il tempo prezioso dei medici. SAM potrebbe essere utilizzato in questa impostazione come strumento per un'annotazione più rapida. Ciò potrebbe essere fatto in diversi modi. Nel caso più semplice, un utente umano fornisce richieste per SAM, che genera una maschera che deve essere approvata o modificata dall'utente; questo potrebbe essere perfezionato in modo iterativo. Un'altra opzione è quella in cui a SAM vengono forniti prompt distribuiti in una griglia sull'immagine (la modalità "segmenta tutto") e genera maschere per più oggetti che vengono quindi nominati, selezionati e/o modificati dall'utente.

SAM che assiste altri modelli di segmentazione. Una versione di questa modalità di utilizzo è quella in cui SAM funziona insieme a un altro algoritmo per segmentare automaticamente le immagini (una "modalità di inferenza"). Ad esempio, SAM, in base a prompt di punti distribuiti nell'immagine, potrebbe generare più maschere di oggetti che potrebbero quindi essere classificati come oggetti specifici da un modello di classificazione separato. Allo stesso modo, un modello di rilevamento indipendente, ad esempio ViTDet, potrebbe generare riquadri di immagini di delimitazione degli oggetti da utilizzare come prompt per SAM per generare maschere di segmentazione precise. Inoltre, SAM potrebbe essere utilizzato nel ciclo di addestramento di qualche altro modello di segmentazione semantica. Ad esempio, le maschere generate da un modello di segmentazione su immagini senza etichetta durante l'addestramento potrebbero essere utilizzate come prompt per SAM per generare maschere più precise per queste immagini, che potrebbero essere utilizzate come esempi di addestramento supervisionato perfezionati in modo iterativo per il modello in fase di addestramento.

Nuovi foundation model della segmentazione per le immagini mediche. In questa modalità di utilizzo, il processo di sviluppo di un nuovo foundation model della segmentazione per le immagini mediche potrebbe essere guidato dal processo di sviluppo proprio di SAM. La difficoltà maggiore risiederebbe nella disponibilità molto inferiore di immagini mediche e annotazioni di qualità rispetto alle immagini naturali, ma in linea di principio ciò è possibile. Un'opzione più fattibile potrebbe essere quella di mettere a punto SAM su immagini e maschere mediche da una varietà di ambiti di imaging medico, piuttosto che addestrarlo da zero, poiché ciò richiederebbe meno immagini.

2.3.3 Metodologia

Nella sezione precedente sono stati descritti vari scenari di utilizzo di SAM per la segmentazione delle immagini mediche. Questi sono concettualmente promettenti ma si basano in gran parte sul presupposto che SAM possa generare segmentazioni accurate di immagini mediche. Di seguito, viene valutata sperimentalmente questa affermazione e le prestazioni di SAM all'interno di una varietà di diversi scenari di utilizzo realistici e dataset nell'imaging medico.

2.3.3.1 Dataset

Hanno compilato e curato una serie di 19 dataset di imaging medico disponibili al pubblico per la segmentazione delle immagini. Il dataset include raggi X planari [33, 34], immagini di risonanza magnetica (MRI) [35, 36, 37, 38, 39], immagini di tomografia computerizzata (CT) [36, 40, 41], immagini ad ultrasuoni (US) [42, 43, 44, 45, 46] e immagini di tomografia a emissione di positroni (PET) [47].

2.3.3.2 Esperimenti

In questi esperimenti hanno eseguito una valutazione approfondita di SAM sia con prompt non iterativi (generati prima dell'applicazione di SAM) sia con prompt iterativi (generati dopo aver visto le previsioni del modello). Hanno anche esplorato la modalità "segmenta tutto" di SAM e analizzato i diversi output che SAM genera in risposta all'ambiguità nei prompt.

Prompt non iterativi. In questa modalit, i prompt sono simulati per riflettere il modo in cui un utente umano potrebbe generarli mentre guarda gli oggetti. Ci si concentra su cinque modalità di prompt non iterativo progettate per catturare i casi di utilizzo realistici di SAM per la generazione di maschere di immagini, utilizzando punti o box di delimitazione. Una cosa essenziale da considerare è che un singolo “oggetto” di interesse/maschera della “ground truth” può essere costituito da più parti sconnesse, il che è particolarmente comune nelle immagini mediche. Le cinque modalità di prompt sono: (1) un punto prompt viene posizionato al centro della regione contigua più grande dell'oggetto di interesse/maschera di verità fondamentale, (2) un punto prompt viene posizionato al centro di ciascuna regione contigua separata dell'oggetto di interesse (fino a tre punti), (3) viene posizionata una box per racchiudere strettamente la regione contigua più grande dell'oggetto di interesse, (4) viene posizionata una box per racchiudere strettamente ciascuna regione contigua separata dell'oggetto di interesse (fino a tre box) e (5) viene posizionata una singola box per racchiudere strettamente l'intera maschera dell'oggetto.

Prompt iterativi. Viene utilizzata una strategia comune e intuitiva per simulare prompt di punti iterativi realistici, che riflette il modo in cui questi potrebbero essere generati da un utente in modo interattivo [48]. Nello specifico, una volta che la rete ha effettuato una previsione, viene calcolata una mappa degli errori in cui sia le previsioni false positive che quelle false negative sono contrassegnate come 1, ovvero il punto più lontano da 0. Quindi si può trovare la posizione del prompt successivo come posizione centrale della componente più grande della maschera di errore.

I prompt possono essere ambigui nel senso che potrebbe non essere chiaro a quale oggetto nell'immagine si riferisca. Uno scenario tipico è quando gli oggetti sono nidificati l'uno nell'altro

nell'immagine. Ad esempio, quando un utente fornisce un prompt relativo a un punto all'interno di una componente necrotica di un tumore al cervello, potrebbe voler segmentare quella componente, l'intero tumore, un emisfero del cervello, l'intero cervello o l'intera testa. In risposta a questo problema, SAM fornisce più output volti a chiarire le ambiguità dei prompt. Questa è una caratteristica molto importante e pratica di SAM poiché nell'impostazione della segmentazione interattiva, all'utente potrebbero essere presentati più output potenziali, dai quali potrà selezionare quello più vicino all'oggetto che intendeva.

In relazione all'ambiguità, in tutti gli esperimenti viene presentata anche ciò che gli sviluppatori di SAM chiamano “prestazioni Oracle”. Questa è la prestazione del modello quando viene sempre utilizzata la previsione più vicina (in termini di IoU) alla maschera reale, ovvero la previsione Oracle, tra le tre previsioni generate da SAM. Quando i prompt vengono generati in modo iterativo, la previsione Oracle viene utilizzata per creare la mappa degli errori che guida la posizione del prompt successivo.

Infine, è stato confrontato SAM con tre metodi di segmentazione interattiva, vale a dire RITM, SimpleClick e FocalClick.

2.3.3.3 Metrica di valutazione delle prestazioni

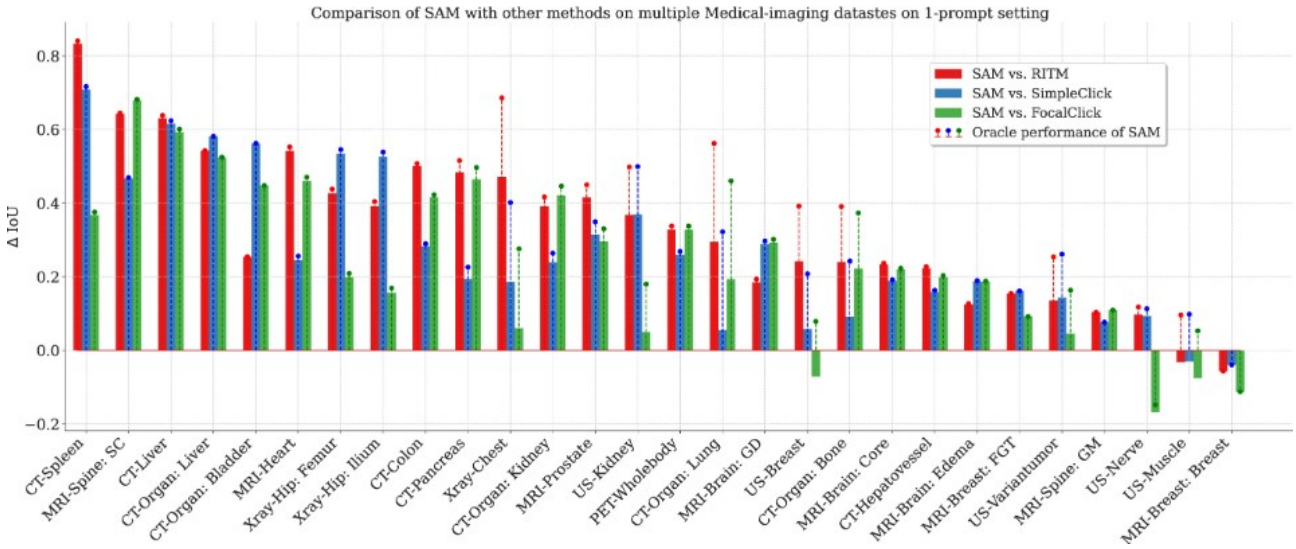
Per ciascun dataset, è stata valutata l'accuratezza delle maschere che SAM genera dati vari prompt, rispetto alle annotazioni della maschera "ground truth" per il dataset e l'attività specificati. Nella valutazione quantitativa, viene utilizzata sempre la maschera con la massima confidenza generata da SAM per un dato prompt. Viene utilizzato IoU come metrica di valutazione, in modo simile al documento originale di SAM. Per analizzare le prestazioni del modello per i dataset contenenti più tipi di oggetti, sono stati calcolati l'IoU per ciascun oggetto separatamente.

2.3.4 Risultati

Dalle prestazioni di SAM per le cinque modalità di utilizzo, vengono tratte diverse conclusioni. Innanzitutto, le prestazioni dei SAM variano ampiamente tra i diversi dataset. Si va da un impressionante IoU di 0.9118 a un IoU molto scarso di 0.1136. Il confronto delle prestazioni per diverse modalità di prompt mostra una chiara superiorità dei prompt box rispetto ai prompt point. Inoltre, come previsto, i prompt in cui ciascuna parte separata dell'oggetto è indicata separatamente sono generalmente superiori a quelli in cui è indicata solo una parte, o tutte le parti sono delineate in una box. Ciò è stato particolarmente pronunciato per i dataset in cui gli oggetti sono generalmente costituiti da più di una parte. Seguendo queste due tendenze, la modalità 4, in cui ciascuna parte di

un oggetto è indicata da una box separata, ha mostrato le migliori prestazioni con un IoU medio di 0.6542.

Inoltre, la modalità Oracle ha mostrato un moderato miglioramento rispetto alla modalità predefinita. L'entità di questo miglioramento dipendeva fortemente dal dataset.



Confronto di SAM con altri tre metodi, ovvero RITM, SimpleClick e Focalclick, con l'impostazione del prompt a 1 punto [2].

Confrontando SAM con altri metodi, si è ottenuto che la prestazione media tra diversi dataset è stata di 0.4595 IoU per SAM, 0.5137 IoU per SAM in modalità Oracle, 0.2240 IoU per FocalClick, 0.1910 IoU per SimpleClick e 0.1322 IoU per RITM. Si noti che SAM mostra prestazioni complessive notevolmente migliori rispetto a tutti gli altri metodi, anche se non viene utilizzato nella modalità più performante di utilizzo delle box come prompt, poiché abbiamo utilizzato prompt puntuali per avere un confronto equo tra i diversi metodi. Se SAM avesse utilizzato la modalità 3 (box singola), avrebbe sovraperformato tutti gli altri metodi in tutte le attività tranne una. La prestazione media per la modalità 3 è stata di 0.5891 IoU. Tuttavia, quando vengono forniti clic aggiuntivi in modo iterativo con l'obiettivo di affinare le segmentazioni restituite dai modelli, la superiorità di SAM diminuisce e viene superata da altri due metodi (SimpleClick e RITM) per cinque o più punti forniti dagli utenti. Ciò è dovuto al fatto che SAM non sembra trarre quasi alcun vantaggio dalle informazioni aggiuntive fornite attraverso i punti interattivi dopo che sono stati forniti due o tre punti. Si vede anche che SAM, quando viene fornito un singolo prompt, ha difficoltà a segmentare oggetti con più regioni non contigue. È più probabile segmentare una regione contigua invece di cercare di trovare ulteriori regioni semanticamente simili nell'intera immagine. Pertanto, nello scenario in cui esistono più regioni di interesse per un oggetto, ulteriori punti prompt per SAM possono essere utili se mirano a

regioni aggiuntive, ma oltre a ciò, il vantaggio di ulteriori punti prompt è trascurabile e in alcuni casi tale input aggiuntivo è dannoso.

2.3.5 Conclusioni

Si è giunti alle seguenti conclusioni.

La precisione di SAM per la segmentazione delle immagini mediche zero-shot è mediamente moderata e varia in modo significativo tra diversi dataset e immagini diverse all'interno di un dataset.

Il modello offre prestazioni migliori con le box prompt, in particolare quando viene fornita una box per ciascuna parte separata dell'oggetto di interesse.

SAM supera RITM, SimpleClick e FocalClick nella stragrande maggioranza delle impostazioni valutate in cui viene fornito un singolo punto di prompt non iterativo.

Nell'impostazione in cui vengono forniti più prompt di punti perfezionati in modo iterativo, SAM ottiene vantaggi molto limitati da prompt di punti aggiuntivi, ad eccezione degli oggetti con più parti. D'altro canto, gli altri algoritmi migliorano notevolmente con ulteriori richieste di punti, fino a superare le prestazioni di SAM. Tuttavia, le modalità di richiesta dei punti sono inferiori alle modalità di richiesta delle box di SAM.

Uno dei contributi di questo studio è che sono state identificate cinque diverse modalità di utilizzo dei metodi di segmentazione interattiva. Ciò è di particolare importanza per i modelli che hanno più componenti, una caratteristica comune nell'imaging medico. Queste modalità hanno mostrato anche prestazioni diverse in tali scenari. Sebbene queste modalità dimostrino la varietà di usi, il lavoro futuro potrebbe concentrarsi sulla prompt engineering, sia non iterativa che iterativa, che potrebbe potenzialmente raggiungere prestazioni ancora più elevate.

Una limitazione notevole di SAM nel contesto dell'imaging medico è che funziona solo in 2D, mentre molte immagini mediche sono 3D. L'algoritmo potrebbe essere esteso al 3D in diversi modi. Un modo semplice sarebbe generare prima la maschera in una determinata sezione in base al prompt fornito e quindi generare automaticamente i prompt nelle sezioni vicine in base alla maschera generata.

Nel complesso, SAM si dimostra promettente per l'uso nelle immagini mediche, a condizione che vengano utilizzate strategie di prompt adeguate per il dataset e il compito scelto. Il lavoro futuro includerà lo sviluppo di diversi modi per adattarlo alla costruzione di modelli specifici per l'imaging medico e all'estensione alla segmentazione 3D.

2.4 SAM per l'analisi delle immagini mediche: 2° esperimento

2.4.1 Introduzione

Sono stati fatti molti esperimenti con SAM sulle immagini mediche. Hanno valutato SAM in modalità Everything segmentando le regioni della lesione in varie strutture anatomiche (ad esempio, cervello, polmone e fegato) e modalità (tomografia computerizzata (CT) e risonanza magnetica (MRI)) [49]. SAM ha dimostrato un'ottima capacità di segmentare organi con contorni ben definiti, tuttavia, la segmentazione di lesioni amorfe ha mostrato risultati meno soddisfacenti. Un altro studio ha poi valutato le prestazioni di SAM in alcuni sottocampi sanitari (segmentazione del disco e della coppa ottica, del polipo e della lesione cutanea) utilizzando sia strategie automatiche Everything che due strategie manuali prompt (punti e box) [22]. I risultati hanno evidenziato la forte dipendenza di SAM dai prompt utente. In particolare, l'utilizzo di punti ha dimostrato di essere cruciale per ottenere segmentazioni accurate, soprattutto in presenza di oggetti complessi o ambigui. Nell'attività di segmentazione del cervello utilizzando la risonanza magnetica [50], hanno confrontato SAM con lo strumento di segmentazione del cervello (BET) della FMRIB Software Library. I risultati quantitativi hanno mostrato che i risultati della segmentazione di SAM erano migliori di quelli di BET, dimostrando il potenziale di SAM per l'applicazione nelle attività di segmentazione del cervello. Hanno, inoltre, valutato le prestazioni di SAM nelle attività di segmentazione della patologia digitale (processo di analisi delle immagini digitali provenienti da tessuti biologici, con l'obiettivo di identificare e delimitare le diverse componenti di interesse), inclusa la segmentazione di tessuti tumorali, non tumorali e nuclei cellulari [16]. I risultati suggeriscono che SAM fornisce risultati di segmentazione eccezionali per oggetti connessi di grandi dimensioni. Tuttavia, potrebbe non raggiungere costantemente prestazioni soddisfacenti per la segmentazione di oggetti con istanze dense, anche con 20 punti per immagine. Infine, hanno applicato SAM all'attività di segmentazione dei polipi utilizzando cinque dataset di riferimento nell'impostazione Everything [51]. I risultati hanno mostrato che, sebbene in alcuni casi SAM possa segmentare accuratamente i polipi, esiste un ampio divario tra SAM e i metodi all'avanguardia.

Più recentemente, diversi studi hanno testato la modalità Everything su 10 dataset o attività MIS (Medical Image Segmentation, si riferisce al processo di segmentazione dell'oggetto desiderato da un'immagine medica 2D o 3D) pubblici [2, 20, 21, 52]. Nel primo studio [20], si è concluso che le prestazioni di segmentazione zero-shot di SAM sono notevolmente inferiori a quelle dei metodi tradizionali basati sul deep learning. Nel secondo [2], gli autori hanno valutato le prestazioni di SAM utilizzando diversi numeri di punti. Hanno osservato che all'aumentare del numero di punti, le prestazioni di SAM convergono. Hanno inoltre notato che le prestazioni di SAM sono (1)

complessivamente moderate e (2) estremamente instabili tra diversi dataset e casi. Nel terzo studio [21] hanno convalidato che SAM originale potrebbe fallire su molti dataset medici con un punteggio DICE medio del 58.52%. Hanno poi perfezionato SAM utilizzando immagini mediche e hanno scoperto che MedSAM proposto ha ottenuto un miglioramento del 22.51% su DICE rispetto al SAM. L'ultimo studio [52] ha adottato la tecnica dell'adattatore per il fine-tuning di SAM e migliorato la sua capacità medica. Gli esperimenti hanno convalidato che l'adattatore SAM medico proposto può superare i metodi MIS all'avanguardia (SOTA) [53].

2.4.2 Metodologia

Sebbene i lavori di cui sopra abbiano studiato le prestazioni di SAM in MIS, presentavano almeno una delle seguenti limitazioni: (1) piccoli dataset, (2) strategia di test SAM unica (diversi oggetti medici spesso presentano caratteristiche diverse e quindi possono avere modalità proprie adatte per i test) e (3) mancanza di valutazioni complete e approfondite (alcuni dei lavori esistenti hanno valutato SAM solo tramite i risultati di visualizzazione forniti dalla demo online).

2.4.2.1 Dataset COSMOS 1050K

Nell'esperimento studiato, per valutare appieno le prestazioni di generalizzazione di SAM in MIS, sono stati raccolti 53 dataset pubblici e sono stati standardizzati per costruire un grande dataset: COSMOS 1050K.

Le immagini mediche coprono un'ampia gamma di tipi di oggetti, come organi cerebrali e tumori [36, 54, 55, 56, 57], polmoni e cuore [33, 58, 59, 60], addome [36, 61, 62, 63, 64, 65], colonna vertebrale [66, 67, 68], cellule [69] e polipi [70, 71]. Per essere compatibili con le diverse modalità di valutazione della SAM, sono stati utilizzati i seguenti criteri di esclusione:

1. Escludere oggetti estremamente piccoli, come la coclea e l'uretere. Ciò è dovuto alla difficoltà di generare automaticamente punti o box prompt su oggetti estremamente piccoli.
2. Eliminare dalla segmentazione gli oggetti nel volume 3D che presentano una significativa discontinuità volumetrica a causa della loro forma anatomica allungata e della procedura di segmentazione sequenziale. Ciò è necessario per prevenire la creazione di bounding box multipli per un singolo oggetto
3. Escludere oggetti con una struttura complessiva relativamente discreta, come immagini istopatologiche del cancro al seno, fette di alberi della trachea, arterie renali e vene. La maggior parte di questi oggetti sono dispersi in più elementi in una sezione 2D e incorporati in altri oggetti, con il risultato che non è possibile utilizzare in modo sensato la modalità prompt di SAM sugli oggetti da verificare.

Secondo i criteri di cui sopra, COSMOS 1050K comprende ora un totale di 84 oggetti.

In totale, COSMOS 1050K è costituito da 1.050.311 immagini o sezioni 2D, di cui 1.003.809 sezioni originate da 8.653 volumi 3D e 46.502 immagini 2D autonome. Inoltre, il dataset incorpora 6.033.198 maschere.

2.4.2.2 Esperimenti

Ricordiamo che nella repository ufficiale GitHub di SAM (https://github.com/facebookresearch/segment-anything), gli autori forniscono tre tipi di modelli preaddestrati con dimensioni backbone diverse, denominati ViT-B, ViT-L e ViT-H. In questo studio, hanno scelto il ViT-B più piccolo (con 12 strati di trasformatore e 91 milioni di parametri) e il ViT-H più grande (con 32 strati di trasformatore e 636 milioni di parametri) come codificatori per eseguire tutte le modalità di test.

Viene dapprima utilizzata la modalità di SAM Everything (S1). La modalità prompt modificata contiene cinque strategie: un punto positivo (S2), cinque punti positivi (S3), cinque punti positivi con cinque punti negativi (S4), una box (S5) e una box con un punto positivo (S6). È stata inoltre stabilita una regola unificata per la selezione dei punti per garantire casualità, ripetibilità e accuratezza. Per la selezione del punto positivo, (a) viene prima calcolato il centro di massa della maschera Ground Truth (GT). (b) Se il centro di massa fosse all'interno della maschera GT, viene preso il centro come primo punto positivo. (c) Quindi, viene appiattita direttamente la maschera GT su un vettore unidimensionale e vengono ottenuti gli altri punti positivi adottando il metodo di campionamento uniforme. (d) Se il centro di massa fosse esterno alla maschera GT, tutti i punti positivi richiesti sarebbero ottenuti eseguendo il passaggio c. Per la selezione dei punti negativi, si è cercato di evitare di selezionare punti troppo distanti dalla regione target. Nello specifico, viene prima ingrandito di due volte il bounding box della GT. I punti negativi sono stati generati anche campionando nella regione non GT. Infine, per la selezione della box, si adotta direttamente il bounding box della maschera GT senza alcuna operazione aggiuntiva. La strategia di cui sopra può garantire la ripetibilità degli esperimenti.

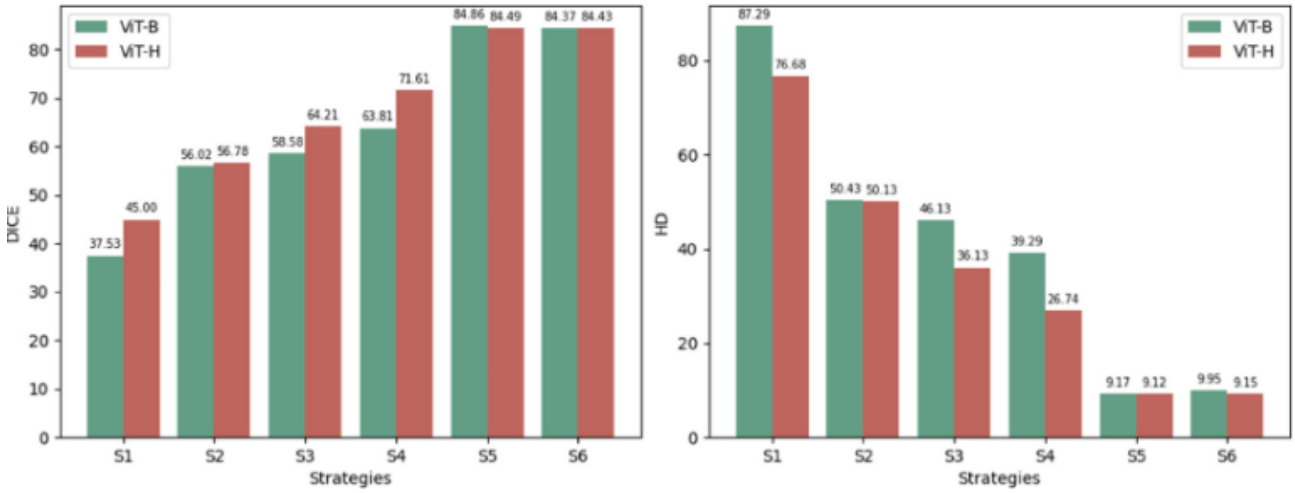
Per velocizzare i test, è stato eseguito l'algoritmo SAM n volte, estraendo ogni volta le caratteristiche principali delle immagini (embedding). Questo processo era molto lento. Per risolverlo, hanno calcolato le caratteristiche una sola volta per ogni immagine e le hanno salvate in file npz, un formato efficiente per salvare array NumPy. In questo modo, per eseguire nuovi test, bastava solo caricare i dati salvati, rendendo il processo molto più veloce (circa n volte). I file npz hanno permesso di salvare

non solo le embedding, ma anche informazioni aggiuntive come i punti e le box necessarie per i test con prompt, ottimizzando ulteriormente il flusso di lavoro.

2.4.2.3 Metrica di valutazione delle prestazioni

Per valutare appieno le prestazioni di segmentazione di SAM, sono stati utilizzati tre parametri: coefficiente DICE [40], coefficiente di somiglianza di Jaccard [72] e Hausdorff Distance [73].

2.4.3 Risultati



Confronto delle prestazioni medie di ViT-B e ViT-H con le diverse strategie (S1-S6) [3].

2.4.3.1 Prestazioni di segmentazione in diversi modelli

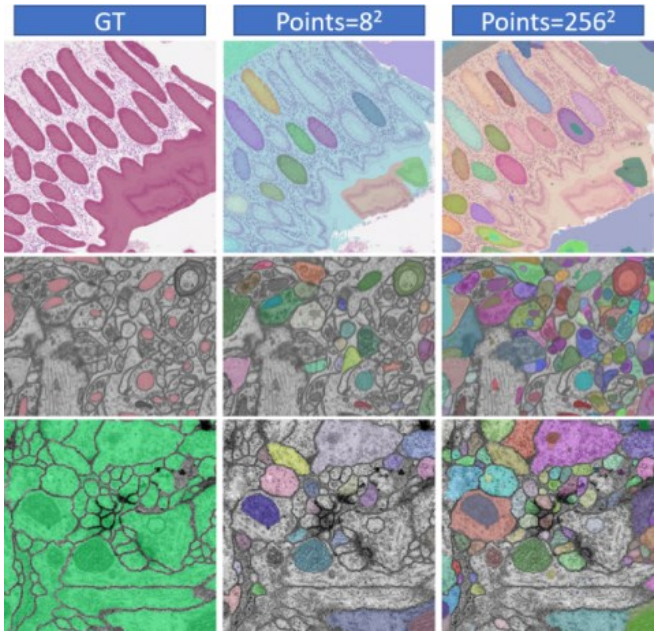
In questa sezione, vengono confrontate le prestazioni di segmentazione tra due modelli (ViT-B e ViT-H) con strategie diverse. Si osserva che, in modalità Everything, ViT-H supera ViT-B di 7.47% su DICE ed è inferiore di 10.61% rispetto a ViT-B su HD. Per il prompt a un punto, ViT-H raggiunge prestazioni medie leggermente superiori a ViT-B. All'aumentare del numero di punti richiesti, i vantaggi di ViT-H diventano più evidenti. Mentre, per le strategie di box senza o con punto, le loro prestazioni sono molto vicine (differenze in DICE: 0.37% e 0.06%). Rispetto al prompt del punto, il prompt della box contiene più informazioni sulla regione dell'oggetto. Pertanto, può guidare meglio SAM con diversi modelli per ottenere migliori prestazioni di segmentazione.

Confrontando la distribuzione DICE per gli stessi oggetti valutati in base a diverse dimensioni del modello, si dimostra che ViT-H mostra prestazioni più stabili con deviazioni standard più piccole rispetto a ViT-B.

2.4.3.2 Prestazioni di segmentazione in diverse modalità di test

In questa sezione, vengono confrontate le prestazioni di segmentazione tra diverse strategie utilizzando diversi modelli (ViT-B e ViT-H). Sia per ViT-B che per ViT-H, gli andamenti prestazionali delle diverse strategie sono sostanzialmente coerenti. La modalità Everything ottiene la prestazione peggiore. Per i prompt point, aggiungendo più punti, quindi passando da S2 a S4, si otterrà un miglioramento stabile delle prestazioni (ViT-B: DICE dal 56.02% al 63.81%, ViT-H: DICE dal 56.78% al 71.61%). SAM con una box fornisce le prestazioni migliori. L'aggiunta di un punto alla box, invece, non porta a miglioramenti significativi. (ViT-B: DICE 0.49%, ViT-H: DICE 0.06%). Sulla base degli esperimenti, si conclude che le box prompt includono maggiori informazioni più vitali rispetto ai point prompt, poiché la box in realtà indica la posizione esatta del bersaglio e anche le sue potenziali caratteristiche di intensità data la regione limitata. Tuttavia, i punti rappresentano solo le caratteristiche parziali del target, il che può creare confusione.

2.4.3.3 Analisi sul numero di punti in modalità Everything



[3] Diversi casi di adenocarcinoma [74], mitocondri [75] e strutture neurali [76].

Come descritto precedentemente, nella modalità Everything, viene visualizzata una griglia di punti (32 x 32). Il numero di punti avrà un impatto sulle prestazioni di segmentazione finale. Soprattutto per le immagini che hanno più target con dimensioni diverse, la progettazione errata dei parametri porterà a una segmentazione imperfetta con alcuni oggetti che non verranno richiesti. Sono stati testati quattro dataset con più oggetti su un'unica immagine. I risultati mostrano che in questi quattro dataset, all'aumentare del numero di punti da 8^2 a 256^2, anche il DICE aumenta gradualmente. Inoltre, viene anche dimostrato che più punti portano più oggetti potenziali. Infine, troppi punti fanno sì che SAM

divida un oggetto in più pezzi, distruggendo l'integrità di questo. Aumentare il numero di punti può portare anche ad un aumento significativo del tempo di test. Si tratta quindi di un compromesso tra prestazioni di segmentazione ed efficienza del test.

2.4.3.4 Tempo di annotazione e analisi della qualità

In questa sezione si discute se SAM può aiutare i medici a migliorare i tempi e la qualità delle annotazioni. Sono state campionate casualmente 100 immagini con prestazioni DICE medie da COSMOS 1050K, per costruire un sottoinsieme di valutazione comprendente 55 oggetti e 620 maschere in 9 modalità, comprese istanze dello stesso oggetto in diverse modalità. Inoltre sono stati invitati tre medici con 10 anni di esperienza per valutare se la previsione di SAM nelle box prompt potesse migliorare la velocità e la qualità dell'annotazione. Sono stati assegnati compiti tra cui (1) annotare da zero tutti gli oggetti nel sottoinsieme di valutazione, (2) modificare le etichette degli oggetti in base alle previsioni di SAM e (3) registrare il tempo per entrambe le attività. Per valutare la qualità delle annotazioni, viene utilizzato l'indice Human Corrrection Efforts (HCE) [77], che stima lo sforzo umano richiesto per correggere previsioni imprecise per soddisfare requisiti specifici di accuratezza (ad esempio, maschere Ground Truth) nelle applicazioni del mondo reale. L'indice HCE inferiore indica che la maschera (annotazione dell'umano con/senza SAM) è più vicina alla Ground Truth, ovvero l'annotazione è di qualità superiore. Si vede che, con l'aiuto di SAM, si può raggiungere una qualità di annotazione più elevata (l'HCE medio passa da 5.07 a 4.80) e aumenta la velocità di annotazione di circa il 25%. Nello specifico, è possibile risparmiare 1.31 minuti per annotare un'immagine (il tempo medio passa da 4.27 minuti a 2.96 minuti) e circa 0.2 minuti per un oggetto (poiché un'immagine contiene circa 6.2 oggetti). Maggiore è il numero di strutture anatomiche da etichettare, più evidente sarà il vantaggio dell'efficienza di SAM.

2.4.3.5 Impatto della diversa casualità dei prompt sulle prestazioni

Negli esperimenti precedenti, sono state testate le prestazioni ottimali teoriche di SAM selezionando il centro di massa e la compacting box perché possono includere le caratteristiche più rappresentative del bersaglio. Tuttavia, non è pratico fare clic sul centro esatto o disegnare il riquadro esatto di ciascun oggetto per valutare SAM. Pertanto, in questa sezione si discute cosa accade aggiungendo diversi livelli di casualità ai centri e alle box per simulare le operazioni umane nella vita reale [78].

Nello specifico, vengono ingranditi/spostati i punti/box in modo casuale in 0-10, 10-20 e 20-30 pixel. Gli esperimenti casuali sono stati condotti tre volte e sono stati calcolati i risultati medi. Il Dice drop rappresenta la diminuzione percentuale del valore DICE medio rispetto ai risultati originali senza spostamento. Per S2 (punto singolo), la performance di DICE è scesa del 2.67%, 7.38% e 14.62%

con l'aumento del livello di spostamento. Con l'aumento del numero di prompt dei punti (S3 e S4), il DICE drop potrebbe essere attenuato, e la stabilità del modello potrebbe essere migliorata. SAM è decisamente peggiorato negli offset delle box (S5, diminuzione delle prestazioni del 24.11% per spostamenti di 20-30 pixel), mentre questo impatto è stato ancora più pronunciato aggiungendo un punto alle box (S6, con un decremento del 29.93%).

2.4.3.6 Confronto tra SAM e metodi interattivi

Nelle sezioni precedenti, vengono inserite tutte le istruzioni una volta nel decodificatore SAM per un confronto equo delle sue prestazioni a un round. Per imitare le procedure di segmentazione interattive della vita reale, è stato eseguito SAM multi-round. La strategia di selezione dei punti è simile ai comuni metodi interattivi. Nello specifico, SAM fa prima clic sul centro del target, quindi i clic rimanenti si basano sulle regioni dei falsi negativi (FN) e dei falsi positivi (FP). Viene quindi confrontato SAM con due diversi approcci di segmentazione interattiva, ovvero FocalClick [79] e SimpleClick [80], entrambi pre-addestrati sullo stesso numero di immagini di SAM.

Sono stati selezionati 10 organi/tumori tipici, coprendo varie modalità, forme, dimensioni e distribuzioni di intensità. Sulla base dei risultati DICE, la conclusione è che: (1) SAM ha sovraperformato FocalClick e SimpleClick nella prima interazione con un singolo punto; (2) Con il progredire delle iterazioni, le prestazioni di SAM aumentavano lentamente, o addirittura diminuivano, mentre le prestazioni dei metodi interattivi potevano essere migliorate costantemente; (3) Utilizzando 10 punti, SAM ha ottenuto risultati peggiori rispetto ai metodi interattivi. Si ritiene, quindi, che l'attuale capacità di iterazione multi-round basata su punti di SAM sia debole sulle immagini mediche.

2.4.3.7 Affinamento specifico dell'attività per SAM

La debole percezione delle capacità di SAM sulla maggior parte delle immagini/attività mediche è dovuta principalmente alla mancanza di dati di addestramento. Il database di addestramento di SAM, ovvero SA-1B (https://ai.meta.com/datasets/segment-anything/), contiene 11 milioni di foto, inclusi luoghi naturali, oggetti e scene, ma senza immagini mediche. Le immagini naturali sono comunemente diverse dalle immagini mediche perché hanno codifica dei colori, definizioni e confini degli oggetti relativamente chiari, primo piano (oggetti) e sfondo (non oggetti) più facili da distinguere e dimensioni relativamente bilanciate. Tuttavia, la maggior parte delle immagini mediche sono in scala di grigi, con confini degli oggetti poco chiari e complessi, sfondo e primo piano simili e dimensioni dell'immagine ad ampio raggio (in particolare contenenti alcuni oggetti molto piccoli).

Pertanto, in questo esperimento hanno messo a punto SAM utilizzando parte del COSMOS 1050K per migliorare la percezione degli oggetti medici da parte di SAM. Nello specifico, sono stati considerati 45 oggetti comuni e tipici per la messa a punto di SAM ed è stata considerata solo la fine-tuning di SAM utilizzando la box prompt. Hanno adattato l'image encoder per ridurre al minimo i costi di calcolo e mantenuto congelato il prompt encoder grazie alla sua potente capacità di codificare le informazioni sulla posizione della box. Pertanto durante la fine-tuning sono stati regolati solo i parametri nel mask decoder.

I risultati dimostrano un miglioramento generale nelle prestazioni di segmentazione dopo la messa a punto per entrambi i modelli ViT-B e ViT-H. Nello specifico, per ViT-B, 32 oggetti su 45 mostrano un miglioramento delle prestazioni, mentre ViT-H mostra un miglioramento in 37 oggetti su 45. Ciò può dimostrare la forte capacità di apprendimento di ViT-H, perché ha quasi 7 volte i parametri di ViT-B (636 milioni contro 91 milioni).

2.4.4 Conclusioni

In questo studio, è stato valutato in modo completo SAM per la segmentazione di un ampio dataset di immagini mediche. Sulla base delle suddette analisi empiriche, le conclusioni sono le seguenti:

1. SAM ha mostrato prestazioni notevoli in alcuni oggetti specifici ma era instabile, imperfetto o addirittura totalmente fallimentare in altre situazioni.
2. SAM con ViT-H ha mostrato prestazioni complessive migliori rispetto a quelle con ViT-B.
3. SAM ha ottenuto risultati migliori con i prompt manuali, in particolare con la box, rispetto alla modalità Everything.
4. SAM potrebbe aiutare l'annotazione umana con un'elevata qualità di etichettatura e meno tempo.
5. SAM è sensibile alla casualità nei prompt del punto centrale e delle box strette e potrebbe subire un grave calo delle prestazioni.
6. SAM ha ottenuto risultati migliori rispetto ai metodi interattivi con uno o pochi punti, ma viene superato all'aumentare del numero di punti.
7. Le prestazioni di SAM sono correlate a diversi fattori, inclusa la complessità dei confini.
8. La fine-tuning di SAM su compiti medici specifici potrebbe migliorare le sue prestazioni.

Infine, si ritiene che, sebbene SAM abbia il potenziale per diventare un modello MIS generale, le sue prestazioni nel compito MIS non sono attualmente stabili.

2.5 Esperimento personale

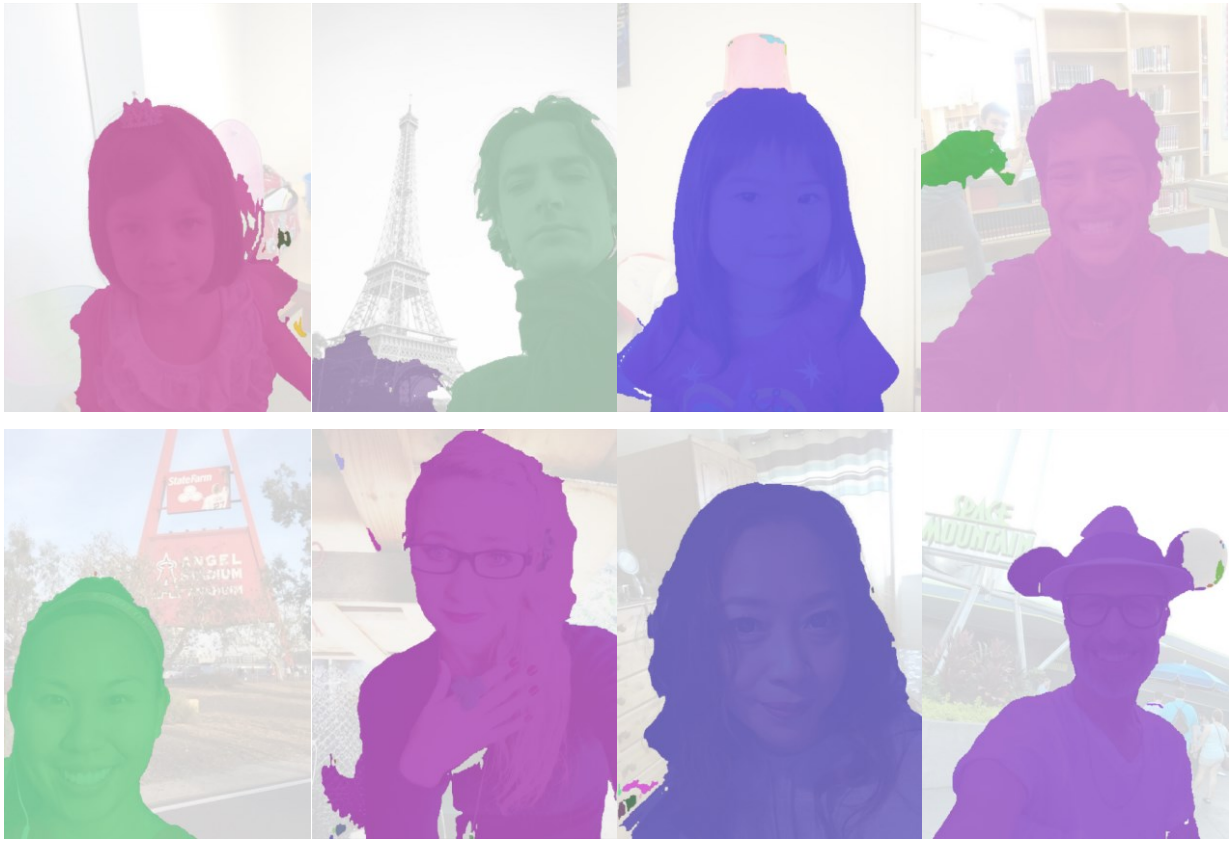
2.5.1 Introduzione

Il modello SAM ha dimostrato di essere uno strumento versatile per la segmentazione di immagini. In particolare, è stato proposto un metodo innovativo [92] per combinare le capacità di SAM con quelle di modelli di segmentazione specializzati.

2.5.2 Metodologia e risultati

L'esperimento da me svolto si concentra nell'apportare modifiche al codice utilizzato nel metodo proposto citato nell'introduzione, in modo da utilizzare tutti i blob nelle maschere di deeplab [93]. I blob sono regioni connesse di pixel che condividono le stesse caratteristiche. In altre parole, un blob è una zona all'interno dell'immagine che rappresenta un oggetto o una parte di un oggetto che è stato identificato dal modello di segmentazione.

Le impostazioni predefinite erano: ViT-H, oracle (in cui le maschere di segmentazione vengono dedotte solo da SAM con checkpoint estratti dalla ground truth) e manual (in versione auto, viene data un'immagine e segmentato l'oggetto ritenuto più rilevante da SAM; in versione manual, viene data una lista di punti e segmentato un oggetto). Ho iniziato confrontando le liste di punti prodotte da tre diversi campionatori sulle maschere DeepLab per il dataset Portrait [94], controllando i punti che erano in una lista e non nell'altra. Successivamente, ho verificato che per il prompt fosse necessario esclusivamente il blob principale. Per farlo, ho controllato, per ciascuna delle 270 immagini, quanti blob venissero generati e le relative quantità di pixel per ciascuno di essi. La maggior parte delle immagini generavano da 1 a 3 blob, di cui uno principale, ovvero il blob con la maggiore area in pixel, e gli altri con poche decine di pixel. C'erano in particolare 8 immagini (43, 96, 97, 113, 160, 196, 223, 230) con cifre diverse dalle altre:



Le immagini sono in ordine numerico da sinistra a destra, nelle quali vengono evidenziati i vari blob segmentati, ciascuno con un colore diverso.

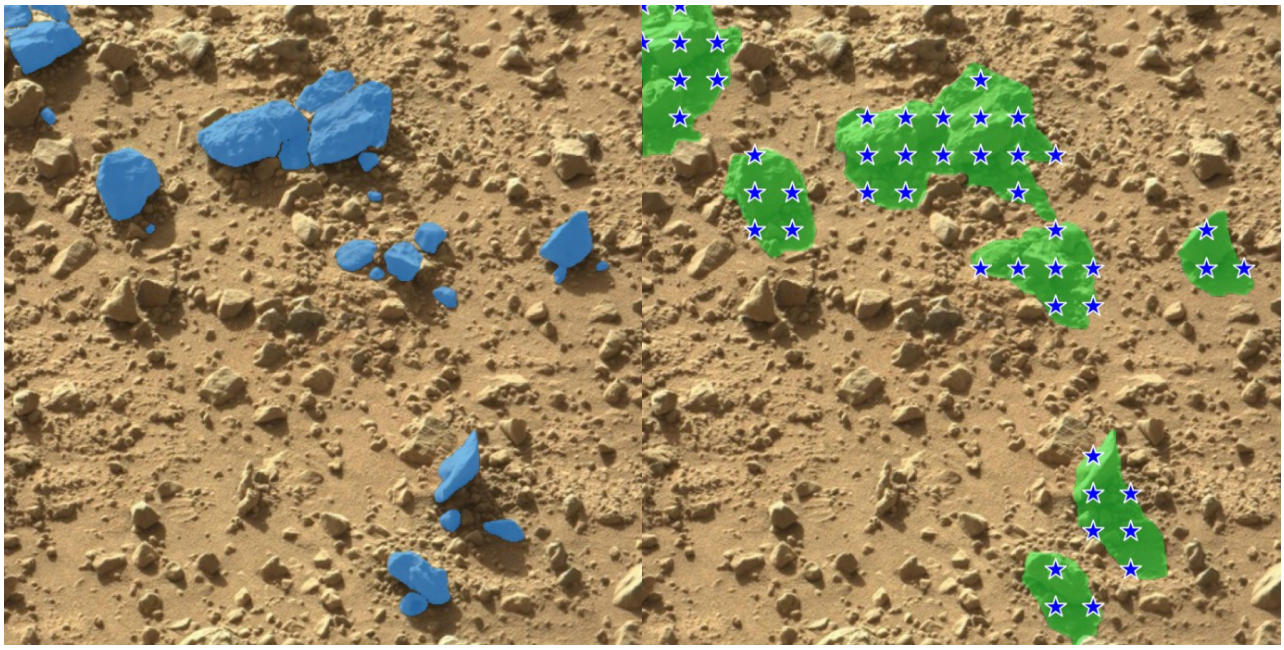
Le immagini sono in ordine da sinistra a destra in ordine numerico.

1. Immagine 43:
    - 1° blob: 202720 pixel;
    - 2° blob: 596 pixel;
    - 3° blob: 47 pixel;
    - 4° blob: 405 pixel.
2. Immagine 96:
    - 1° blob: 145341 pixel;
    - 2° blob: 28467 pixel.
3. Immagine 97:
    - 1° blob: 412 pixel;
    - 2° blob: 36 pixel;
    - 3° blob: 42 pixel;
    - 4° blob: 233650 pixel;
    - 5° blob: 361 pixel.

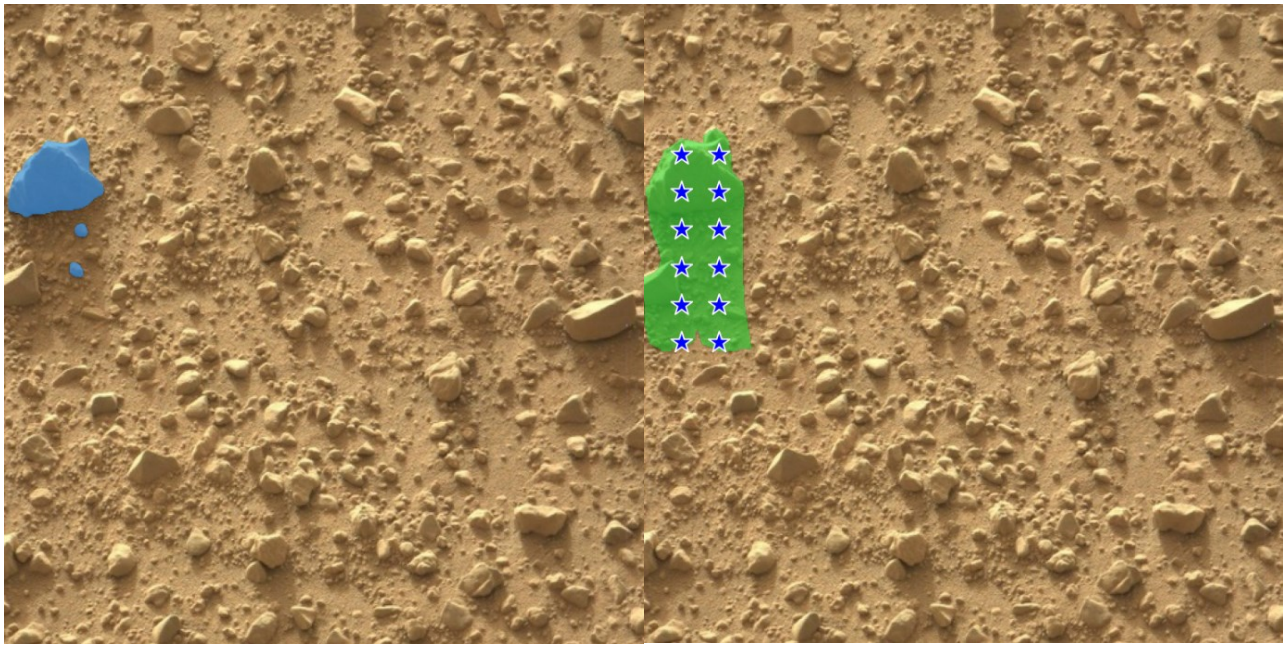
4. Immagine 113:
    - 1° blob: 242101 pixel;
    - 2° blob: 13649 pixel.
5. Immagine 160:
    - 1° blob: 32 pixel;
    - 2° blob: 136127.
6. Immagine 196:
    - 1° blob: 302522 pixel;
    - 2° blob: 314 pixel;
    - 3° blob: 26 pixel;
    - 4° blob: 362 pixel;
    - 5° blob: 24 pixel;
    - 6° blob: 327 pixel;
    - 7° blob: 25 pixel;
    - 8° blob: 49 pixel;
    - 9° blob: 58 pixel.
7. Immagine 223:
    - 1° blob: 309222 pixel;
    - 2° blob: 45 pixel;
    - 3° blob: 833 pixel;
    - 4° blob: 76 pixel;
    - 5° blob: 30 pixel;
    - 6° blob: 428 pixel;
    - 7° blob: 354 pixel.
8. Immagine 230:
    - 1° blob: 191837 pixel;
    - 2° blob: 23 pixel;
    - 3° blob: 23 pixel;
    - 4° blob: 470 pixel;
    - 5° blob: 59 pixel;
    - 6° blob: 111 pixel.

Dalle immagini visualizzate, si vede che i blob aggiuntivi rispetto a quello principale sono errati o, nella migliore delle ipotesi, irrilevanti per produrre il prompt e possono essere ignorati.

Sono successivamente passato al dataset Mars [95] e a SAM in modalità “auto”. Le maschere DeepLab che ci sono per questo dataset non hanno blob multipli, quindi il problema precedente non si pone. Si pone però per l’output, dato che la funzione di generazione restituisce una lista di record. Ora, quindi, l’obiettivo è capire quale euristica di selezione delle maschere sperimentare. Di conseguenza, ho iniziato visualizzando, per ciascuna immagine del dataset, l’immagine e i relativi punteggi Dice e IoU.



Rispettivamente i blob rilevati da SAM sull’immagine 155 e la ground truth dell’immagine 155.



Rispettivamente i blob rilevati da SAM sull’immagine 158 e la ground truth dell’immagine 158.

I dati ottenuti segnalano 7 immagini (su 167) su cui SAM ottiene un basso IoU e anche un basso Dice score: immagine 14 (IoU: 37,6%; Dice: 54,67%), immagine 29 (IoU: 43,2%; Dice: 60,37%), immagine 34 (IoU: 39,9%; Dice: 57,04%), immagine 66 (IoU: 49,8%; Dice: 66,46%), immagine 90 (IoU: 32,7%; Dice: 49,28%), immagine 155 (IoU: 47,9%; Dice: 64,78%) e immagine 158 (IoU: 28,9%; Dice: 44,82%). Tuttavia, nelle immagini 155 e 158 la ground truth è sbagliata perché include l'ombra delle rocce.

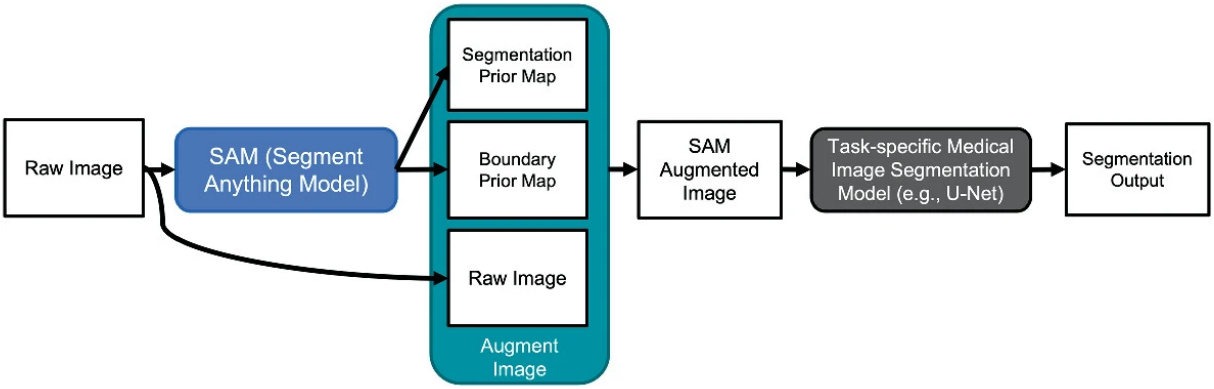
Infine ho visualizzato per le precedenti immagini e per altre immagini con solamente punteggio IoU basso, tutti i blob, compresi quelli scartati dall'euristica, generati da SAM. Da questi, si nota che l'euristica scarta lo sfondo, che è corretto, e negli altri casi non scarta niente di erroneo e, quindi, anche qui il suo comportamento è corretto.

Qui il mio esperimento si conclude perché non riuscivo a scrivere il codice necessario per continuarlo.

Capitolo 3. Input Augmentation con SAM

3.1 Introduzione

Nonostante la forte capacità di SAM nel produrre segmentazione per un'ampia varietà di oggetti, diversi studi (come i due esperimenti precedenti) hanno dimostrato che non è abbastanza potente per attività di segmentazione che richiedono conoscenze specialistiche nel settore (ad esempio segmentazione di immagini mediche).



Panoramica dell'input augmentation con SAM per potenziare la segmentazione delle immagini mediche [4].

Input augmentation (aumento degli input) è una tecnica utilizzata per migliorare le prestazioni di un modello di machine learning. Nel caso specifico di questo studio [4], consiste nell'aggiungere informazioni aggiuntive alle immagini mediche grezze prima di sottoporle al modello di segmentazione. Queste informazioni aggiuntive sono rappresentate dalle maschere di segmentazione generate da SAM.

In sostanza, l'input augmentation permette di:

- Arricchire i dati: Le maschere di segmentazione forniscono informazioni più dettagliate sulle regioni di interesse nelle immagini mediche, aiutando il modello a focalizzare l'attenzione sulle aree più importanti.
- Migliorare le prestazioni: L'aggiunta di queste informazioni extra può aiutare il modello a generalizzare meglio e a ottenere risultati più accurati.
- Sfruttare conoscenze preesistenti: SAM, essendo pre-addestrato su un vasto dataset, porta con sé conoscenze generali sulla segmentazione che possono essere utili per il compito specifico delle immagini mediche.

Viene proposto un nuovo metodo chiamato SAMAug che funziona come segue:

1. Generazione delle maschere: SAM viene utilizzato per generare maschere di segmentazione per le immagini mediche.
2. Fusione delle informazioni: Le maschere generate vengono combinate con le immagini originali attraverso una funzione di fusione, creando così un nuovo input arricchito.
3. Addestramento del modello: Il modello di segmentazione delle immagini mediche viene addestrato su questi nuovi input aumentati.

3.2 Metodologia

3.2.1 Segmentazione e mappe a priori dei confini

Nell'impostazione del prompt a griglia, SAM genera sistematicamente maschere di segmentazione su tutta l'immagine. Queste maschere, ciascuna con un associato punteggio di stabilità, vengono poi accumulate in una mappa di segmentazione a priori, dove la trasparenza di ogni maschera riflette il suo livello di affidabilità. Parallelamente, viene creata una mappa dei confini, delineando i contorni esterni di tutte le maschere generate. Queste mappe forniscono una rappresentazione completa delle possibili segmentazioni dell'immagine.

3.2.2 Aumentare le immagini di input

L'aumento dell'immagine avviene concatenando l'immagine originale in scala di grigi con le mappe a priori di segmentazione e dei confini lungo la dimensione dei canali. In questo modo, l'immagine di input viene estesa con informazioni semantiche sulla segmentazione e sui contorni degli oggetti, creando una rappresentazione più completa e informativa per il modello.

3.2.3 Addestramento del modello con immagini SAM-Augmented

Applicando l'aumento dell'input a ogni immagine del set di addestramento, hanno ottenuto un nuovo dataset arricchito di informazioni semantiche. Questo dataset viene quindi utilizzato per addestrare direttamente un modello di segmentazione medica standard. L'obiettivo dell'addestramento è di far apprendere al modello a prevedere le segmentazioni corrette a partire dalle immagini aumentate, che incorporano le mappe di segmentazione e dei confini generate da SAM. In questo modo, il modello impara a sfruttare le conoscenze pre-addestrate di SAM per migliorare le proprie prestazioni.

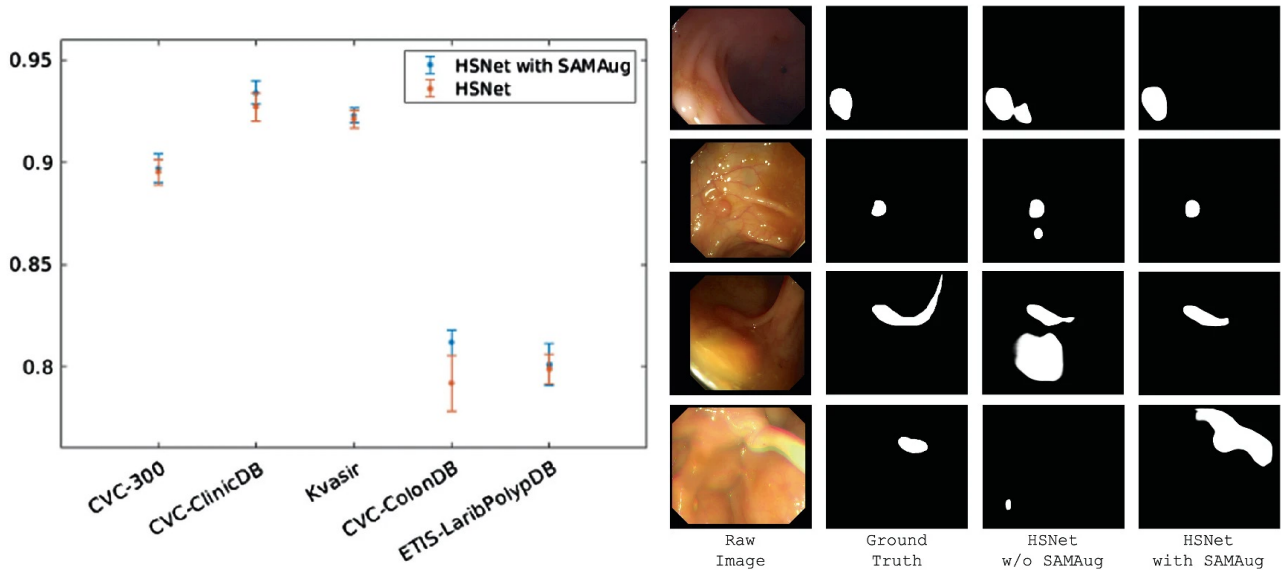
3.3 Esperimenti e risultati

3.3.1 Dataset e configurazioni

Per valutare l'efficacia del metodo SAMAug, vengono eseguiti esperimenti sui benchmark Polyp [51], MoNuSeg [81] e GlaS [74]. Per gli esperimenti di segmentazione dei polipi, viene seguita la configurazione di addestramento utilizzata nell'addestramento del modello all'avanguardia (SOTA) HSNet [82]. Per la segmentazione MoNuSeg e GlaS, l'addestramento di un modello di segmentazione di immagini mediche utilizza l'ottimizzatore Adam [83], con dimensione batch uguale a 8, dimensione finestra di ritaglio immagine uguale a 256×256, e tasso di apprendimento uguale a 5 e-4. Il numero totale di iterazioni di training è 50K.

Per le valutazioni, vengono utilizzati il DICE score, l'F-score (valuta le prestazioni di segmentazione cellulare a livello di pixel) e l'AJI (Aggregated Jaccard Index), una metrica specifica per la valutazione della segmentazione di oggetti multipli

3.3.2 Segmentazione dei polipi su cinque dataset

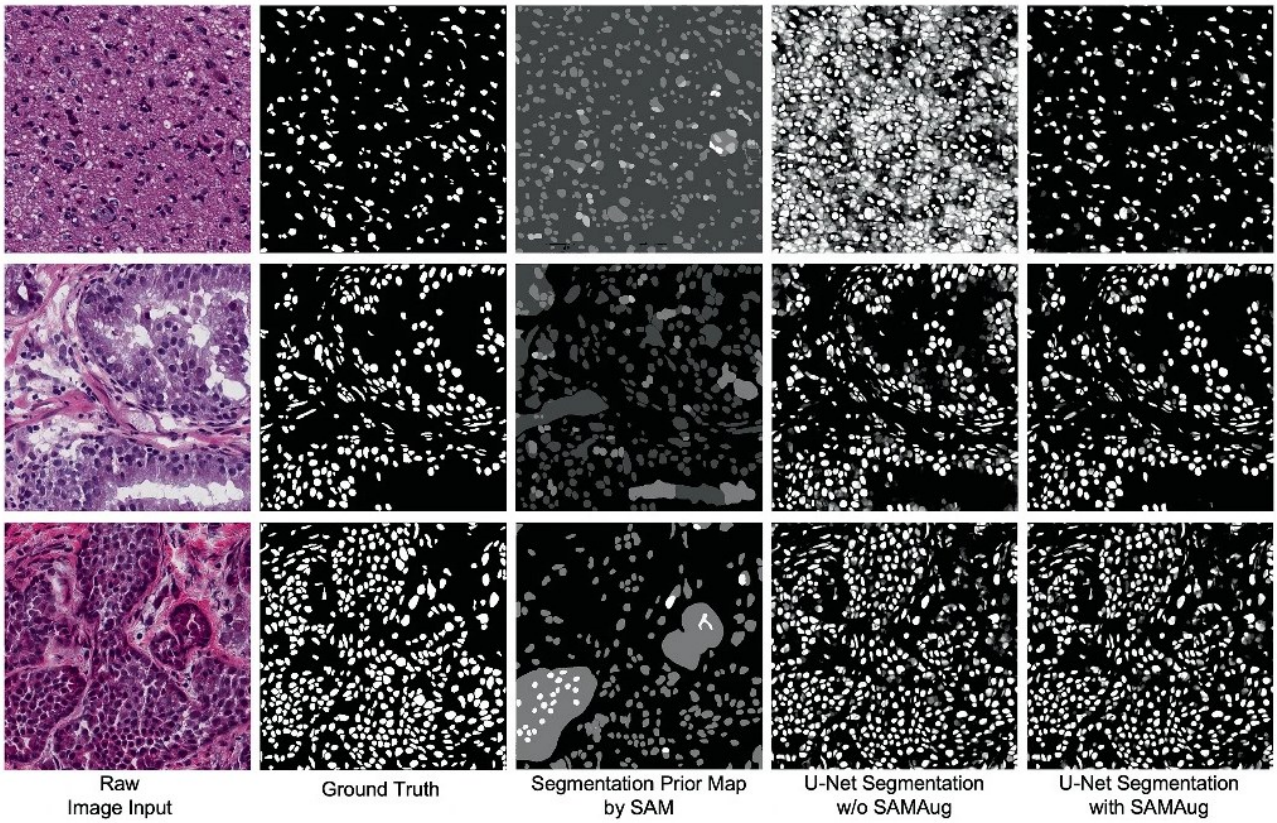


A sinistra sono riportati i risultati numerici della segmentazione dei polipi di HSNet base e HSNet potenziato con SAMAug. A destra, invece, sono riportati i corrispondenti risultati visivi [4].

Tutte le sessioni di addestramento del modello HSNet sono state eseguite dieci volte con diversi seeds casuali per riportare le medie e le deviazioni standard delle prestazioni di segmentazione. Si osserva che SAMAug migliora significativamente HSNet sui dataset CVC-ClinicDB [84] e CVC-ColonDB [85] e rimane allo stesso livello di prestazioni sugli altri tre dataset (Kvasir [86], CVC-300 [87] e ETIS-LaribPolypDB [88]).

3.3.3 Segmentazione cellulare sul dataset MoNuSeg

Il set di addestramento MoNuSeg è composto da 30 immagini con circa 22000 annotazioni nucleari cellulari. Il set di test contiene 14 immagini con circa 7000 annotazioni nucleari cellulari. Queste annotazioni, create manualmente, indicano la posizione esatta e la forma di ogni singolo nucleo cellulare nelle immagini.



Confronti visivi dei risultati della segmentazione sul dataset MoNuSeg [4].

Si notano chiari vantaggi del metodo proposto, SAMAug, nel migliorare i risultati di segmentazione per i modelli U-Net [89] (senza SAMAug si ha AJI: 58.36% e F-score: 75.70%; con SAMAug si ha AJI: 64.30% e F-score: 82.36%), P-Net [90] (senza SAMAug si ha AJI: 59.46% e F-score: 77.09%; con SAMAug si ha AJI: 63.98% e F-score: 82.56%) e Attention U-Net [91] (senza SAMAug si ha AJI: 58.76% e F-score: 75.43%; con SAMAug si ha AJI: 63.15% e F-score: 81.49%). Si nota anche che, sebbene la segmentazione generata da SAM non fornisca immediatamente una segmentazione cellulare accurata, SAM fornisce una segmentazione percettiva generale prima che i successivi modelli DL generino risultati di segmentazione specifica per attività molto più accurate.

3.3.4 Segmentazione delle ghiandole nel dataset Glas

Il dataset GlaS contiene 85 immagini di training (37 benigne (BN), 48 maligne (MT)) e 60 immagini di test (33 BN, 27 MT) nella parte A e 20 immagini di test (4 BN, 16 MT) nella parte parte B. Per semplicità, viene unita la parte A del set di test e la parte B del set di test ed eseguita la valutazione della segmentazione contemporaneamente per tutti i campioni nel set di test. Si nota che U-Net con l'aumento SAMAug funziona notevolmente meglio (F-score: 82.50%, DICE score: 87.44%) di quello senza l'aumento SAMAug (F-score: 79.33%, DICE score: 86.35%).

3.4 Conclusione

Esperimenti su tre compiti di segmentazione hanno mostrato l'efficacia del metodo SAMAug. Il lavoro futuro potrebbe prendere in considerazione la possibilità di condurre ulteriori ricerche su: (1) progettazione di una funzione di aumento più robusta e avanzata; (2) migliorare l'efficienza dell'applicazione di SAM nello schema SAMAug; (3) utilizzo di SAMAug per stime di incertezza e in altre applicazioni orientate alla clinica.

Riferimenti

1. Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, Ross Girshick. Segment Anything. (2023) arXiv:2304.02643.

2. Mazurowski M.A., Dong H., Gu H., Yang J., Konz N., Zhang Y. Segment anything model for medical image analysis: an experimental study. https://doi.org/10.1016/j.media.2023.102918. Med. Image Anal., 89 (2023), Article 102918.

3. Yuhao Huang, Xin Yang, Lian Liu, Han Zhou, Ao Chang, Xinrui Zhou, Rusi Chen, Junxuan Yu, Ji ongquan Chen, Chaoyu Chen, Sijing Liu, Haozhe Chi, Xindi Hu, Kejuan Yue, Lei Li, Vicente Grau, Deng-Ping Fan, Fajin Dong, Dong Ni. Segment anything model for medical images? https://doi.org/10.1016/j.media.2023.103061.

4. Yizhe Zhang, Tao Zhou, Shuo Wang, Peixian Liang, Yejia Zhang, Danny Z. Chen. Input Augmentation with SAM: Boosting Medical Image Segmentation with Segmentation Foundation Model. https://doi.org/10.1007/978-3-031-47401-9_13.

5. Xian Lin, Yangyang Xiang, Li Yu, Zengqiang Yan. SAMUS: Adapting Segment Anything Model for Clinically-Friendly and Generalizable Ultrasound Image Segmentation. (2023) arXiv:2309.06824.

6. Jiaqi Wang, Zhengliang Liu, Lin Zhao, Zihao Wu, Chong Ma, Sigang Yu, Haixing Dai, Yiheng Liu, Songyao Zhang, Enze Shi, Yi Pan, Tuo Zhang, Dajiang Zhu, Xiang Li, Xi Jiang, Bao Ge, Yixuan Yuan, Dinggang Shen, Tianming Liu, Shu Zhang. Review of large vision models and visual prompt engineering. https://doi.org/10.1016/j.metrad.2023.100047.

7. Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. CVPR, 2019.

8. Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. IJCV, 2020.

9. Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ADE20K dataset. IJCV, 2019.

10. Candice Schumann, Susanna Ricco, Utsav Prabhu, Vittorio Ferrari, and Caroline Pantofaru. A step toward more inclusive people annotations for fairness. Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, 2021.

11. David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. ICCV, 2001.

12. Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra ´ Malik. Contour detection and hierarchical image segmentation. TPAMI, 2010.

13. Yanghao Li, Hanzi Mao, Ross Girshick, and Kaiming He. Exploring plain vision transformer backbones for object detection. ECCV, 2022.

14. Z. Liu, X. Yu, L. Zhang, et al. Deid-gpt: zero-shot medical text de-identification by gpt-4. arXiv preprint arXiv:2303.11032 (2023).

15. A. Ramesh, M. Pavlov, G. Goh, et al. Zero-shot text-to-image generation. International Conference on Machine Learning, PMLR (2021), pp. 8821-8831.

16. R. Deng, C. Cui, Q. Liu, et al. Segment anything model (sam) for digital pathology: assess zero-shot segmentation on whole slide imaging. Med Imag Deep Learn Short paper Track (2023).

17. P. Shi, J. Qiu, S.M.D. Abaxi, H. Wei, F.P.W. Lo, W. Yuan. Generalist vision foundation models for medical imaging: a case study of segment anything model on zero-shot medical segmentation. Diagnostics, 13 (11) (2023), p. 1947.

18. S. Roy, T. Wald, G. Koehler, et al. Sam. md: zero-shot medical image segmentation capabilities of the segment anything model. arXiv preprint arXiv:2304.05396 (2023).

19. C. He, K. Li, Y. Zhang, et al. Weakly-supervised concealed object segmentation with sam-based pseudo labeling and multi-scale feature grouping. arXiv preprint arXiv:2305.11003 (2023).

20. S. He, R. Bao, J. Li, P.E. Grant, Y. Ou. Accuracy of segment-anything model (SAM) in medical image segmentation tasks. arXiv preprint arXiv:2304.09324 (2023).

21. J. Ma, B. Wang. Segment anything in medical images. arXiv preprint arXiv:2304.12306 (2023).

22. W. Ji, J. Li, Q. Bi, W. Li, L. Cheng. Segment anything is not always perfect: an investigation of SAM on different real-world applications. arXiv preprint arXiv:2304 (2023), Article 05750.

23. Z. Ma, X. Hong, Q. Shangguan. Can sam count anything? an empirical study on sam counting. arXiv preprint arXiv:2304.10817 (2023).

24. S. Julka, M. Granitzer. Knowledge distillation with segment anything (SAM) model for planetary geological mapping. arXiv preprint arXiv:2305.07586 (2023).

25. J. Zhang, Z. Zhou, G. Mai, L. Mu, M. Hu, S. Li. Text2seg: remote sensing image semantic segmentation via text-guided visual foundation models. arXiv preprint arXiv:2304.10597 (2023).

26 D. Wang, J. Zhang, B. Du, D. Tao, L. Zhang. Scaling-up remote sensing segmentation dataset with segment anything model. Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2023).

27. T. Chen, L. Zhu, C. Ding, et al. Sam fails to segment anything?–sam-adapter: adapting sam in underperformed scenes: camouflage, shadow, and more. arXiv preprint arXiv:2304.09148 (2023).

28. H. He, J. Zhang, M. Xu, J. Liu, B. Du, D. Tao. Scalable mask annotation for video text spotting. arXiv preprint arXiv:2305.01443 (2023).

29. Y. Zhang, R. Jiao. How segment anything model (SAM) boost medical image segmentation? arXiv preprint arXiv:2305.03678 (2023).

30. X. Zhou, X. Li, K. Hu, Y. Zhang, Z. Chen, X. Gao. Erv-net: an efficient 3d residual neural network for brain tumor segmentation. Expert Syst Appl, 170 (2021), Article 114566.

31. G. Lu, S. Li, G. Mai, et al. Agi for agriculture. arXiv preprint arXiv:2304.06136 (2023).

32. X. Yang, H. Dai, Z. Wu, et al. Sam for poultry science. arXiv preprint arXiv:2305.10254 (2023).

33. Jaeger S., Candemir S., Antani S., Wáng Y.-X.J., Lu P.-X., Thoma G. Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. Quant. Imaging Med. Surg., 4 (6) (2014), p. 475.

34. Gut D. X-ray images of the hip joints, 1. (2021), 10.17632/zm6bxzhmfz.1.

35. Prados F., Ashburner J., Blaiotta C., Brosch T., Carballido-Gamio J., Cardoso M. J., Conrad B. N., Datta E., Dávid G., De Leener B., et al. Spinal cord grey matter segmentation challenge. Neuroimage, 152 (2017), pp. 312-329.

36. Simpson A.L., Antonelli M., Bakas S., Bilello M., Farahani K., Van Ginneken B., Kopp-Schneider A., Landman B.A., Litjens G., Menze B., et al. A large annotated medical image dataset

for the development and evaluation of segmentation algorithms. (2019) arXiv preprint arXiv:1902.09063.

37. Lemaître G., Martí R., Freixenet J., Vilanova J.C., Walker P.M., Meriaudeau F. Computer-aided detection and diagnosis for prostate cancer based on mono and multi-parametric MRI: A review. Comput. Biol. Med., 60 (2015), pp. 8-31.

38. Menze B.H., Jakab A., Bauer S., Kalpathy-Cramer J., Farahani K., Kirby J., Burren Y., Porz N., Slotboom J., Wiest R., et al. The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans. Med. Imaging, 34 (10) (2014), pp. 1993-2024.

39. Saha A., Harowicz M.R., Grimm L.J., Kim C.E., Ghate S.V., Walsh R., Mazurowski M.A. A machine learning approach to radiogenomics of breast cancer: A study of 922 subjects and 529 DCE-MRI features. Brit. J. Cancer, 119 (4) (2018), pp. 508-516.

40. Bilic P., Christ P., Li H.B., Vorontsov E., Ben-Cohen A., Kaissis G., Szeskin A., Jacobs C., Mamani G. E. H., Chartrand G., et al. The liver tumor segmentation benchmark (LITS). Med. Image Anal., 84 (2023), Article 102680.

41. Rister B., Shivakumar K., Nobashi T., Rubin D.L. CT-ORG: A Dataset of CT Volumes With Multiple Organ Segmentations. The Cancer Imaging. Archive (2019), 20.7937/TCIA.2019.TT7F4V70. URL https://wiki.cancerimagingarchive.net/x/OgWkAw, Version Number: 1 Type: dataset.

42. Al-Dhabyani W., Gomaa M., Khaled H., Fahmy A. Dataset of breast ultrasound images. Data Brief, 28 (2020), Article 104863.

43. Song Y., Zheng J., Lei L., Ni Z., Zhao B., Hu Y. CT2US: Cross-modal transfer learning for kidney segmentation in ultrasound images with synthesized data. Ultrasonics, 122 (2022), Article 106706.

44. Marzola F., van Alfen N., Doorduin J., Meiburger K.M. Deep learning segmentation of transverse musculoskeletal ultrasound images for neuromuscular disease assessment. Comput. Biol. Med., 135 (2021), Article 104623.

45. Anna M., Hasnin M., kaggle446 H., shirzad A., Will C., yffud. Ultrasound Nerve Segmentation

Kaggle (2016) URL https://kaggle.com/competitions/ultrasound-nerve-segmentation.

46. Zhao Q., Lyu S., Bai W., Cai L., Liu B., Wu M., Sang X., Yang M., Chen L. A multi-modality ovarian tumor ultrasound image dataset for unsupervised cross-domain semantic segmentation. (2022) arXiv preprint arXiv:2207.06799.

47. Gatidis S., Hepp T., Früh M., La Fougère C., Nikolaou K., Pfannender C., Schölkopf B., Küstner T., Cyran C., Rubin D. A whole-body FDG-PET/CT dataset with manually annotated tumor lesions. Sci. Data, 9 (1) (2022), p. 601.

48. Mahadevan, S., Voigtlaender, P., Leibe, B., 2018. Iteratively Trained Interactive Segmentation. In: British Machine Vision Conference.

49. Ji G.-P., Fan D.-P., Xu P., Cheng M.-M., Zhou B., Van Gool L. SAM struggles in concealed scenes–empirical study on "segment anything". (2023) arXiv preprint arXiv:2304.06022.

50. Mohapatra S., Gosai A., Schlaug G. Brain extraction comparing segment anything model (SAM) and FSL brain extraction tool. (2023) arXiv preprint arXiv:2304.04738.

51. Zhou T., Zhang Y., Zhou Y., Wu Y., Gong C. Can SAM segment polyps? (2023) arXiv preprint arXiv:2304.07583.

52. Wu J., Fu R., Fang H., Liu Y., Wang Z., Xu Y., Jin Y., Arbel T. Medical SAM adapter: Adapting segment anything model for medical image segmentation. (2023) arXiv:2304.12620.

53. Isensee F., Jaeger P.F., Kohl S.A., Petersen J., Maier-Hein K.H. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods, 18 (2) (2021), pp. 203-211.

54. Shapey J., Wang G., Dorent R., Dimitriadis A., Li W., Paddick I., Kitchen N., Bisdas S., Saeed S. R., Ourselin S., et al. An artificial intelligence framework for automatic segmentation and volumetry of vestibular schwannomas from contrast-enhanced T1-weighted and high-resolution T2-weighted MRI. J. Neurosurg., 134 (1) (2019), pp. 171-179.

55. Bakas S., Reyes M., Jakab A., Bauer S., Rempfler M., Crimi A., Shinohara R. T., Berger C., Ha S. M., Rozycki M., et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge. (2018) arXiv preprint arXiv:1811.02629.

56. Sun Y., Gao K., Wu Z., Li G., Zong X., Lei Z., Wei Y., Ma J., Yang X., Feng X., et al. Multi-site infant brain segmentation algorithms: the iSeg-2019 challenge. IEEE Trans. Med. Imaging, 40 (5) (2021), pp. 1363-1376.

57. Podobnik G., Strojan P., Peterlin P., Ibragimov B., Vrtovec T. HaN-Seg: The head and neck organ-at-risk CT & MR segmentation dataset. Med. Phys. (2023).

58. Bernard O., Lalande A., Zotti C., Cervenansky F., Yang X., Heng P.-A., Cetin I., Lekadir K., Camara O., Ballester M. A. G., et al. Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? IEEE Trans. Med. Imaging, 37 (11) (2018), pp. 2514-2525.

59. Viniavskyi O., Dobko M., Dobosevych O. Weakly-supervised segmentation for disease localization in chest x-ray images. Artificial Intelligence in Medicine: 18th International Conference on Artificial Intelligence in Medicine, AIME 2020, Minneapolis, MN, USA, August 25–28, 2020, Proceedings 18, Springer (2020), pp. 249-259.

60. Setio A.A.A., Traverso A., De Bel T., Berens M.S., Van Den Bogaard C., Cerello P., Chen H., Dou Q., Fantacci M. E., Geurts B., et al.Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge. Med. Image Anal., 42 (2017), pp. 1-13.

61. Zhou, Y., Li, Z., Bai, S., Wang, C., Chen, X., Han, M., Fishman, E., Yuille, A.L., 2019b. Prior-aware neural network for partially-supervised multi-organ segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10672–10681.

62. Zhao Z., Chen H., Wang L. A coarse-to-fine framework for the 2021 kidney and kidney tumor segmentation challenge. Kidney and Kidney Tumor Segmentation: MICCAI 2021 Challenge, KiTS 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceeding, Springer (2022), pp 53-58.

63. Ma J., Zhang Y., Gu S., Zhu C., Ge C., Zhang Y., An X., Wang C., Wang Q., Liu X., Cao S., Zhang Q., Liu S., Wang Y., Li Y., He J., Yang X. AbdomenCT-1K: Is abdominal organ segmentation a solved problem? IEEE Trans. Pattern Anal. Mach. Intell., 44 (10) (2022), pp. 6695-6714, 10.1109/TPAMI.2021.3100536.

64. Ji Y., Bai H., Yang J., Ge C., Zhu Y., Zhang R., Li Z., Zhang L., Ma W., Wan X., et al. AMOS: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. (2022) arXiv preprint arXiv:2206.08023.

65. Luo X., Liao W., Xiao J., Chen J., Song T., Zhang X., Li K., Metaxas D. N., Wang G., Zhang S. WORD: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from CT image. Med. Image Anal., 82 (2022), Article 102642.

66. Sekuboyina A., Husseini M. E., Bayat A., Löffler M., Liebl H., Li H., Tetteh G., Kukačka J., Payer C., Štern D., et al. Verse: A vertebrae labelling and segmentation benchmark for multi-detector CT images. Med. Image Anal., 73 (2021), Article 102166.
67. Löffler M.T., Sekuboyina A., Jacob A., Grau A.-L., Scharr A., El Husseini M., Kallweit M., Zimmer C., Baum T., Kirschke J. S. A vertebral segmentation dataset with fracture grading. Radiol.: Artif. Intell., 2 (4) (2020), Article e190138.

68. Pang S., Pang C., Zhao L., Chen Y., Su Z., Zhou Y., Huang M., Yang W., Lu H., Feng Q. SpineParseNet: spine parsing for volumetric MR image by a two-stage segmentation framework with semantic image representation. IEEE Trans. Med. Imaging, 40 (1) (2020), pp. 262-273.

69. Lee G., Kim S., Kim J., Yun S.-Y. MEDIAR: Harmony of data-centric and model-centric for multi-modality microscopy. (2022) arXiv:2212.03465.

70. Jha D., Smedsrud P.H., Riegler M.A., Halvorsen P., de Lange T., Johansen D., Johansen H.D. Kvasir-seg: A segmented polyp dataset. International Conference on Multimedia Modeling, Springer (2020), pp. 451-462.

71. Hicks S.A., Jha D., Thambawita V., Halvorsen P., Hammer H.L., Riegler M.A. The EndoTect 2020 challenge: evaluation and comparison of classification, segmentation and inference time for endoscopy. Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10-15, 2021, Proceedings, Part VIII, Springer (2021), pp. 263-274.

72. Crum W.R., Camara O., Hill D.L. Generalized overlap measures for evaluation and validation in medical image analysis. IEEE Trans. Med. Imaging, 25 (11) (2006), pp. 1451-1461.

73. Zhou D., Fang J., Song X., Guan C., Yin J., Dai Y., Yang R. Iou loss for 2d/3d object detection. 2019 International Conference on 3D Vision (3DV), IEEE (2019), pp. 85-94.

74. Sirinukunwattana K., Pluim J.P., Chen H., Qi X., Heng P.-A., Guo Y. B., Wang L. Y., Matuszewski B. J., Bruni E., Sanchez U., et al. Gland segmentation in colon histology images: The glas challenge contest. Med. Image Anal., 35 (2017), pp. 489-502.

75. Lucchi, A., Li, Y., Fua, P., 2013. Learning for structured prediction using approximate subgradient descent with working sets. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1987–1994.
76. Cardona A., Saalfeld S., Preibisch S., Schmid B., Cheng A., Puloskas J., Tomancak P., Hartenstein V. An integrated micro-and macroarchitectural analysis of the drosophila brain by computer-assisted serial section electron microscopy. PLoS Biol., 8 (10) (2010), Article e1000502.

77. Qin X., Dai H., Hu X., Fan D.-P., Shao L., Van Gool L. Highly accurate dichotomous image segmentation. European Conference on Computer Vision, Springer (2022), pp. 38-56.

78. Huang Y., yang X., Zou Y., Chen C., Wang J., Dou H., Ravikumar N., Frangi A. F., Zhou J., Ni D. Flip learning: Erase to segment. Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part I 24, Springer (2021), pp. 493-502.

79. Chen, X., Zhao, Z., Zhang, Y., Duan, M., Qi, D., Zhao, H., 2022. Focalclick: Towards practical interactive image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1300–1309.

80. Liu Q., Xu Z., Bertasius G., Niethammer M. SimpleClick: Interactive image segmentation with simple vision transformers. (2022) arXiv preprint arXiv:2210.11006.

81. Kumar, N., Verma, R., Sharma, S., Bhargava, S., Vahadane, A., Sethi, A.: A dataset and a technique for generalized nuclear segmentation for computational pathology. IEEE Trans. Med. Imaging 36(7), 1550–1560 (2017).

82. Zhang, W., Chong, F., Zheng, Yu., Zhang, F., Zhao, Y., Sham, C.-W.: HSNet: a hybrid semantic network for polyp segmentation. Comput. Biol. Med. 150, 106173 (2022).

83. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014).

84. Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F.: WM-DOVA maps for accurate polyp highlighting in colonoscopy: validation vs. saliency maps from physicians. Comput. Med. Imaging Graph. 43, 99–111 (2015).

85. Tajbakhsh, N., Gurudu, S.R., Liang, J.: Automated polyp detection in colonoscopy videos using shape and context information. IEEE Trans. Med. Imaging 35(2), 630–644 (2015).

86. Jha, D., et al.: Kvasir-SEG: a segmented polyp dataset. In: Ro, Y.M., et al. (eds.) MMM 2020. LNCS, vol. 11962, pp. 451–462. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-37734-2_37.

87. Vázquez, D., et al.: A benchmark for endoluminal scene segmentation of colonoscopy images. J. Healthcare Eng. 2017, 1–9 (2017).

88. Silva, J., Histace, A., Romain, O., Dray, X., Granado, B.: Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer. Int. J. Comput. Assist. Radiol. Surg. 9, 283–293 (2014).

89. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28.
90. Wang, G., et al.: DeepIGeoS: a deep interactive geodesic framework for medical image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 41(7), 1559–1572 (2018).
91. Oktay, O., et al. Attention U-Net: learning where to look for the pancreas. In: International Conference on Medical Imaging with Deep Learning (2018).
92. Loris Nanni, Daniele Fusaro, Carlo Fantozzi e Alberto Pretto. Improving Existing Segmentators Performance with Zero-Shot Segmentators. Entropy 2023, 25, 1502.
93 Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation.
In Proceedings of the 15th European Conference on Computer Vision—ECCV, Munich, Germany, 8–14 September 2018; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 833–851.
94. Le, T.N.; Nguyen, T.V.; Nie, Z.; Tran, M.T.; Sugimoto, A. Anabranch Network for Camouflaged Object Segmentation. Comput. Vis. Image Underst. 2019, 184, 45–56.
95. Haiqiang Liu, Meibao Yao, Xueming Xiao, Yonggang Xiong. RockFormer: A U-Shaped Transformer Network for Martian Rock Segmentation. IEEE Transaction on Geoscience and Remote Sensing (Volume: 61).