# SAMのプロンプトチューニングと繰り返し推論による細胞画像セグメンテーションの高精度化

○舩井祥吾 †，　平川翼 †，山下隆義 †，藤吉 弘亘 †

○ Shogo FUNAI †　Tsubasa HIRAKAWA †，Takayoshi YAMASHITA †，and　Hironobu FUJIYOSHI †

†：中部大学，　{funai@mprg.cs, hirakawa@mprg.cs, takayoshi@isc, fujiyoshi@isc}.chubu.ac.jp

＜要約＞セグメンテーションの基盤モデルである Segment Anything Model (SAM) のプロンプトチューニングと，繰り返し推論によりセグメンテーションする手法を提案する．提案手法は，SAM の Prompt encoder 部に新たなトークンを追加してプロンプトチューニングを行うことと，繰り返し回数 $t$ と $t-1$ のマスクを比べて，しきい値よりも大きい変化がある場合に繰り返し推論を行うことにより，位置ずれが生じたプロンプトによる精度低下を防ぐことが可能になる．電子顕微鏡細胞画像を対象とした ISBI データセットと Electron Microscopy Dataset を用いた評価実験により，提案手法の有効性を確認した．

＜キーワード＞電子顕微鏡細胞画像，セグメンテーション，プロンプトチューニング

## 1　はじめに

セグメンテーションとは，ピクセルレベルでのクラス識別を行うことで，オブジェクトの形状やクラスを識別することができる．医療画像や細胞画像のセグメンテーションは，生物学的に関連する形態学的情報をピクセルレベルで識別することができるため，正確な診断や治療計画のサポートに用いられる．しかし，細胞形状には様々な形状があり，学習が困難である．

Segment Anything Model (SAM) [1] は，セグメンテーションタスクの基盤モデルである．プロンプトに基づいて前景と背景にセグメンテーションを行う．例えば，プロンプトにバウンディングボックス (bbox) を使用する場合，bbox で囲んだ領域をセグメンテーションすることができる．このように，クラスに縛られることなく様々な形状に対応が可能である．一方で，プロンプトは人の手によって与えられる．SAM はプロンプトに基づいてセグメンテーションを行うため，プロンプトに位置ずれが生じた場合，精度の低下する．

そこで本研究では，SAM に位置ずれを許容する新たなプロンプトチューニングと，繰り返し推論による最適化を導入し，SAM によるセグメンテーション精度の向上を図る．評価実験により，bbox の位置ずれによる精度の低下を抑制できることを確認した．

## 2　関連研究

Mazurowski らは，SAM のプロンプトとセグメンテーション能力に関する詳細を明らかにした [5]．主な結果は以下の通りである．1 つ目は，セグメンテーションの精度にはデータセットごとに差がある．2 つ目は，プロンプトが正確である場合は適切なセグメンテーションが得られる．3 つ目は，bbox をプロンプトとした場合 point をプロンプトにした場合よりも高精度なセグメンテーション結果が得られる．4 つ目は，複数の point をプロンプトに利用した場合，セグメンテーションの精度がわずかに向上する傾向がある．これらの結果から，SAM は与えるプロンプトが重要である．また，ロボティクス分野においても point と比べて bbox を用いた場合に性能が高いことが示されており [11]，より適切なプロンプトの入力が必要である．

SAM は，様々な画像に対して高い汎用性を持っている一方で，SAM の学習用データに含まれていない顕微鏡画像や細胞画像といった医療用画像に対するセグメンテーション性能は不十分である．このため，電子顕微鏡細胞画像や医療用画像を用いて SAM をファインチューニングする様々な手法が提案されている [4, 6,

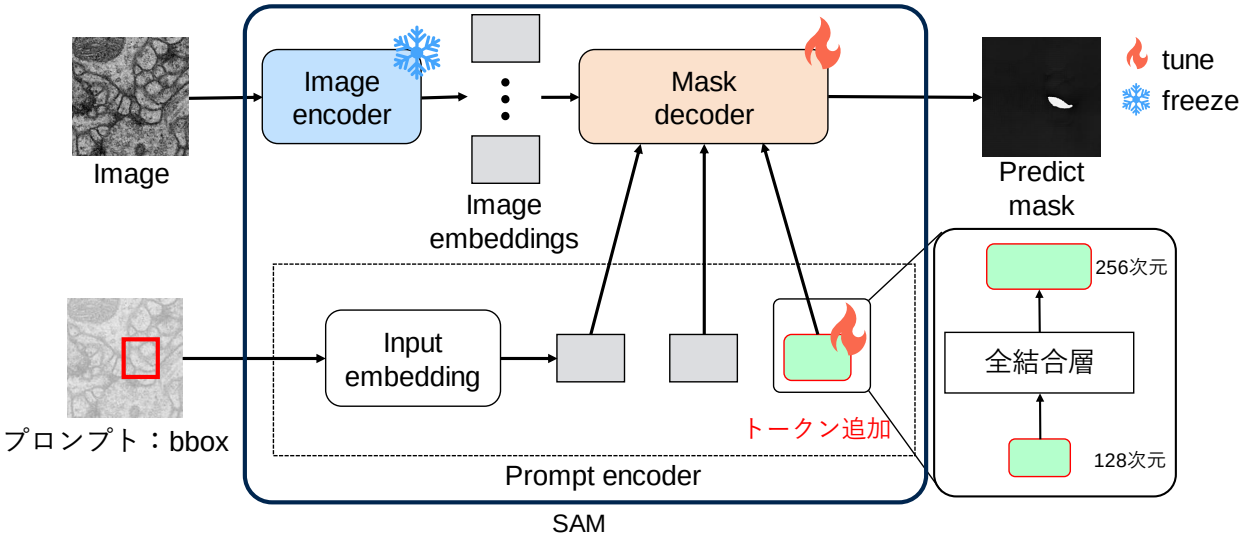


**図 1　トークンの追加とプロンプトチューニング**

7, 9, 10, 12]．MedSAM [12] は，CT，MRI，病理画像などの 1,090,486 枚の画像とマスクのペアから構成されている大規模な医療用画像データセットを用いて SAM のファインチューニングが行われた．MedSAM では Image encoder と Mask decoder をファインチューニングし，プロンプトとして bbox が使用された．bbox は通常の SAM と同じ 0〜20pixel の位置ずれで学習を行っている．そのため，20pixel 程度の位置ずれであれば，SAM とに対する頑健性を獲得している．しかし，学習時以上の位置ずれが発生した場合，精度が低下する可能性がある．

## 3　提案手法

本研究では，位置ずれを起こした bbox に対して，適切なセグメンテーションを出力するために，新たな学習パラメータであるプロンプトの導入と，繰り返し推論を提案する．

### 3.1　プロンプトチューニング

SAM には既にプロンプトチューニング用の学習パラメータが存在しており，SAM 独自の学習パラメータ (Learned embedding) は Positional encoding を行った埋め込みに加算している．一般的なプロンプトチューニングでは埋め込みに加算するのではなく，トークンとして与えられる．そこで，提案手法では位置ずれを考慮するための学習パラメータを新たに追加する．学習パラメータは，トークンとして SAM の Prompt encoder に追加し，プロンプトチューニングを行う．追加したトークンを用いたプロンプトチューニングを図 1 に示す．また，このトークンを全結合層により 128 次元から 256 次元に変換することで，トークンはより複雑な表現を可能にする．

### 3.2　繰り返し推論

推論時に位置ずれした不正確な bbox によるプロンプトを用いた場合，良いマスクを出力することは難しい．そこで推論時に，繰り返し推論を行う．繰り返し推論の流れを図 2 に示す．繰り返し推論は以下のステップで行われる．

**step1** 入力画像とプロンプトを用いて推論を行う．プロンプトにはランダムな位置ずれを含む bbox を用いる．

**step2** 推論した Predict Mask から外接矩形を bbox として設定する．

**step3** 入力画像と設定した bbox を用いて再度推論を行う．

**step4** 推論した $M_t$ と 1 回目に推論した $M_{t-1}$ を比較する．繰り返し推論の条件式を式 (1) に示す．ここで，$M_{t-1}$ は一つ前の Predict mask，$M_t$ は現在の Predict mask，$t$ は現在の繰り返し回数，$HW$ は画像サイズ，$\tau$ はしきい値である．

$$\sum^{HW} |M_t - M_{t-1}| > \tau \tag{1}$$

**step5** 比較した結果

表 1　ISBI データセットの実験結果

| 手法 | prompt tuning | 繰り返し推論 | bbox の位置ずれ | | | | |
|---|---|---|---|---|---|---|---|
| | | | 0% | 10% | 20% | 30% | 40% |
| SAM (FT) | | | 0.822 | 0.810 | 0.774 | 0.699 | 0.604 |
| SAM | ✓ | | 0.825 | 0.810 | 0.777 | 0.702 | 0.606 |
| 提案手法 | ✓ | | **0.830** | **0.820** | 0.785 | 0.714 | 0.622 |
| 提案手法 | ✓ | ✓ | 0.830 | 0.817 | **0.784** | **0.727** | **0.650** |

表 2　Electron Microscopy Dataset の実験結果

| 手法 | prompt tuning | 繰り返し推論 | bbox の位置ずれ | | | | |
|---|---|---|---|---|---|---|---|
| | | | 0% | 10% | 20% | 30% | 40% |
| SAM (FT) | | | **0.885** | **0.875** | 0.844 | 0.772 | 0.673 |
| SAM | ✓ | | **0.885** | 0.874 | 0.844 | 0.772 | 0.673 |
| 提案手法 | ✓ | | 0.884 | 0.874 | 0.845 | 0.779 | 0.683 |
| 提案手法 | ✓ | ✓ | 0.882 | 0.867 | **0.855** | **0.824** | **0.762** |

**しきい値を超える場合** step2 に戻る

**しきい値を超えない場合** predict mask をマスクとして出力する．

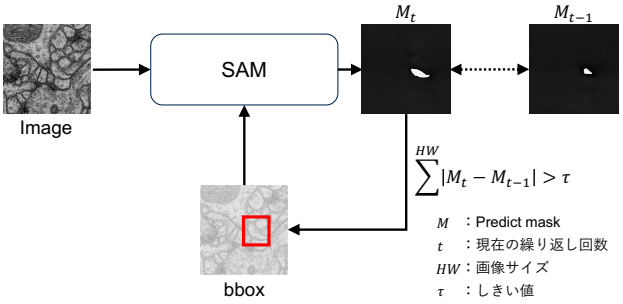


図 2　繰り返し推論

## 4　評価実験

本実験では提案手法の有効性を示すために，トークンのプロンプトチューニングと繰り返し推論による評価実験を行う．ベースラインは，SAM のファインチューニングとする．SAM のファインチューニングでは，Image encoder を固定し，Mask decoder のみを学習する．0%から 40%の位置をずらした bbox をプロンプトとして入力した場合のセグメンテーションを行う．プロンプトチューニングでは Image encoder と Learned embedding のパラメータは固定し，追加したプロンプトと Mask decoder を様々な bbox で学習する．また，通常プロンプトチューニングを行う場合は，既存のモデルのパラメータを凍結させるが，SAM は Mask decoder の学習を行うことで精度が向上することが分かっており，Image encoder よりも Mask decoder は軽量でパラメータ数が少ないため同時に学習を行う．SAM のプロンプトチューニングでは，Prompt encoder の Learned embedding と Mask decoder を学習させる．通常の SAM のファインチューニングでは，Mask decoder のみを学習する．

### 4.1　データセット

本実験では，電子顕微鏡細胞画像を用いた細胞のセグメンテーションのデータセットである ISBI [2] と，Electron Microscopy Dataset [3] を用いる．ISBI は RGB 画像とセグメンテーション画像がペアの 30 枚の画像であり，画像枚数が少ない．そこで 1 枚の画像を 4 つに分割し，120 枚にして使用する．Electron Microscopy Dataset は 3 次元画像である．3 次元画像は 2 次元画像の積み重ねであるため，2 次元画像を 1 枚ずつ取り出すことで変換を行い，173 枚の RGB 画像とセグメンテーション画像のペアを使用する．これらのデータセットは，bbox が用意されていないため，Ground truth のセグメンテーションの外接矩形を bbox の真値のプロンプトとして使用する．ISBI ではデータセット全体の 60%を学習用，20%を検証用，20%をテスト用に使用する．Electron Microscopy Dataset は学習用とテスト用があるため，テスト用の半分を検証用に使用する．

### 4.2　実験概要

プロンプトチューニングでは，60%の確率で 0%，20%の確率で± 10%，± 20%のサイズをずらした bbox をプロンプトとして入力し学習を行う．位置ずれに対する評価では，0%から 40%のサイズをずらした bbox をプロンプトとして入力した場合の Predict mask と Ground truth 間の mIoU を比較する．繰り返し推論ではしきい

値を 2pixel とする．

### 4.3 実験結果

ISBI データセットでの実験結果を表 1 に，Electron Microscopy Dataset での実験結果を表 2 に示す．提案手法は両方のデータセットにおいて 30%から 40%と位置ずれが大きい場合に SAM を超える認識性能を達成した．プロンプトチューニングのみを導入した場合，提案手法が両方のデータセットにおいて，位置ずれが 30%以上で SAM を超える認識性能を達成していることから，プロンプトチューニング時に学習していない位置ずれ度合いに対してある程度頑健になったと言える．また，プロンプトチューニングに加えて繰り返し推論を導入することで，さらに認識性能が改善することから，繰り返し推論が有効であると言える．

### 4.4 マスクの可視化

ISBI データセットのそれぞれの手法を可視化した Predict mask (青色領域)，Ground truth を図 3 に示す．真値の bbox を入力した推論では，SAM (FT) よりもトークンのプロンプトチューニングを行うことで，より True Positive の部分のセグメンテーションをすることができた．位置ずれを起こした bbox では，SAM(FT) と SAM のプロンプトチューニングでは大きな変化は見受けられなかったが，トークンの導入により，マスクの False Positive が少なくなり，より精密な検出ができるようになった．

### 4.5 繰り返し過程の可視化

ISBI データセットでの繰り返し推論における 1 回目から 3 回目の Predict mask (青色領域)，Ground truth を図 4 に示す．小さい bbox が入力された推論では，推論を繰り返すことで Predict mask が Ground truth へ近づいていくものが多く存在した．大きな bbox では，多くの場合，2 度目の推論で良い結果となった．トークンを追加したプロンプトチューニングでは 1 度目の推論でより輪郭を捉えられていることから，bbox の位置ずれが小さくなりやすく，精度が向上したと考えられる．この結果より，繰り返し推論が有効であると言える．bbox が繰り返し推論する際に，もともと小さい物体の場合，bbox が元よりも小さくなってしまった場合に bbox が小さくなり続けることによる精度低下が見受けられた．その結果，bbox の位置ずれが 20%までは学習したことにより，もともと高いセグメンテーション結果が返されるが，小さい物体の繰り返し推論により，良くなる bbox よりも悪くなる bbox の方が多いと考えられる．繰り返し推論の位置ずれを起こした大きい bbox では，正解とは異なるマスクをセグメンテーションしてしまう場合で精度が低下した．

## 5 おわりに

本研究では SAM における位置ずれのある不正確なプロンプトによる性能低下を改善するプロンプトチューニングと繰り返し推論を提案した．トークンのプロンプトチューニングは位置ずれした bbox に対して精度低下を低減した．繰り返し推論では位置ずれが 30%以上で大きく精度低下を低減した．繰り返し推論は 10%から 20%で繰り返し推論はうまくいかなかった．今後は，マスクから作成した bbox に対する損失設計によるプロンプトの最適化を行う．

## 参考文献


[1] A. Kirillov, *et al.*, “Segment anything”, ICCV, 2023.

[2] A. Cardona, *et al.*,“An integrated micro-and macroararchitectural analysis of the drosophila brain by computer-assisted serial section electron microscopy”, PLOS Biology, vol.8, no.10, pp.1-17, 2010.

[3] A. Lucchi, *et al.*, “Supervoxel-based segmentation of mitochondria in EM image stacks with learned shape features”, IEEE Trans Med Imaging, vol.31, no.2, pp.474-486, 2012.

[4] J. Wu, *et al.*, “Medical SAM Adapter: Adapting Segment Anything Model for Medical Image Segmentation”, arXiv preprint arXiv:2304.12620, 2023.

[5] M. Mazurowski, *et al.*, “Segment Anything Model for Medical Image Analysis: an Experimental Study” , arXiv preprint arXiv:2304.10517, 2023.

[6] H. Dai, *et al.*, "SAMAug: Point Prompt Augmentation for Segment Anything Model", arXiv preprint arXiv:2307.01187,2023.

[7] Z. Qiu, *et al.*, "Learnable Ophthalmology SAM", arXiv preprint arXiv:2304.13425,2023.

[8] M. Jia, *et al.*, "Visual Prompt Tuning", arXiv preprint arXiv:2203.12119, 2022.

[9] T. Chen, *et al.*, "SAM Fails to Segment Anything?", arXiv preprint arXiv:2304.09148, 2023.

[10] R. Deng, *et al.*, "Segment Anything Model (SAM) for Digital Pathology: Assess Zero-shot Segmentation on Whole Slide Imaging", arXiv preprint arXiv:2304.04155, 2023.

[11] An Wang, *et al.*,"SAM Meets Robotic Surgery: An Empirical Study in Robustness Perspective", arXiv preprint arXiv:2304.14674, 2023.

[12] Jun Ma, *et al.*,"Segment Anything in Medical Images", arXiv preprint arXiv:2304.12306.

[13] Tao Yu, *et al.*,"Inpaint Anything: Segment Anything Meets Image Inpainting", arXiv preprint arXiv:2304.06790, 2023

[14] Siyuan Huang, *et al.*,"Instruct2Act: Mapping Multi-modality Instructions to Robotic Actions with Large Language Model", arXiv preprint arXiv:2305.11176, 2023.

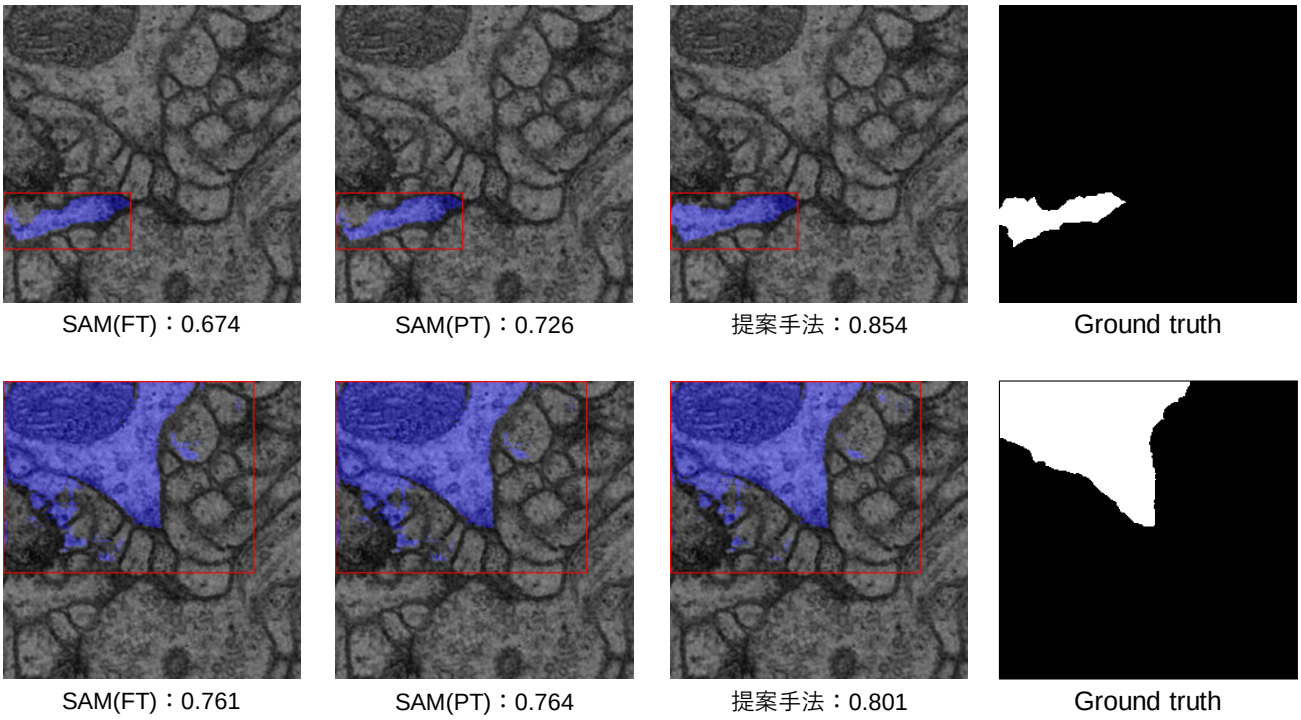


**図 3　ISBI データセットの可視化**

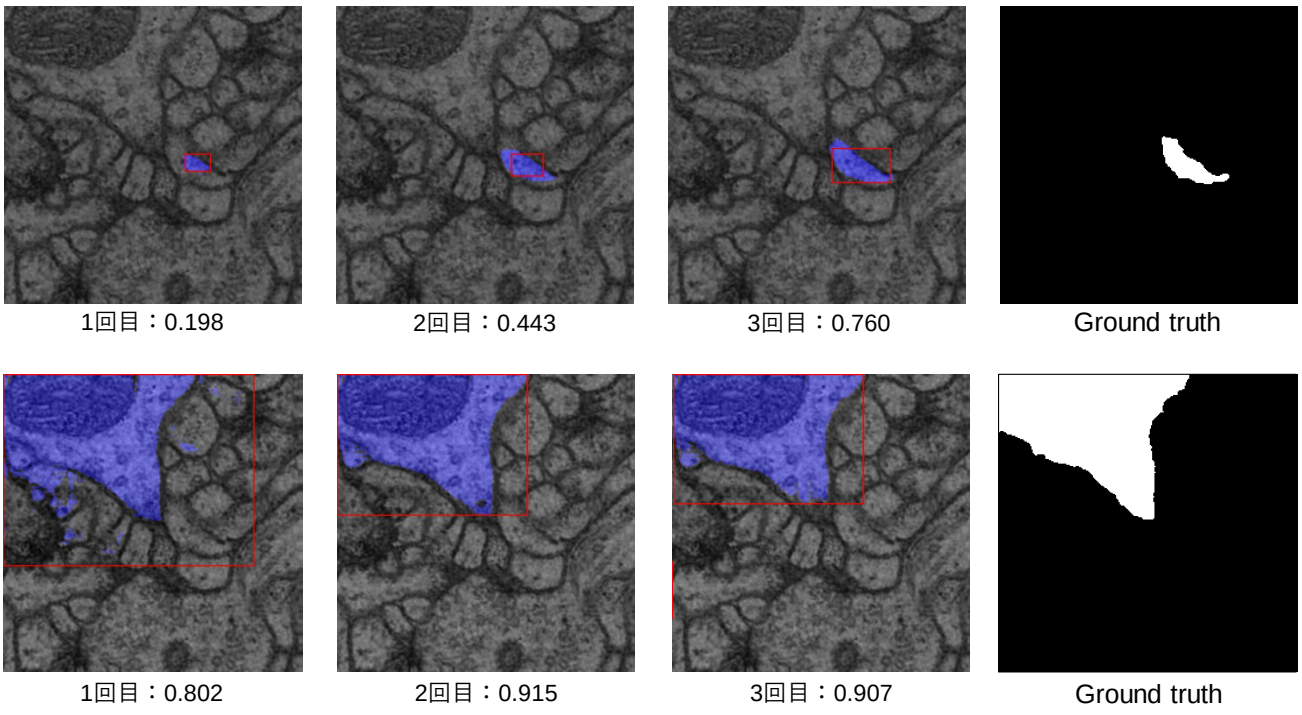


**図 4　ISBI データセットの繰り返し推論による Predict mask の変化**