jh250050

# シミュレーションを活用した手術シーン認識に関する研究

小田昌宏（名古屋大学）

**概要**

外科手術におけるコンピュータ支援は、対象の複雑さから未だ十分実現されていない。今後必要となる自律的手術支援ロボット開発を見据え、本研究では画像ベースの高精度な手術シーン認識の実現を目指す。複雑な術中画像の認識理解を実現するため、大規模データからの手術シーン用画像基盤モデルを構築した。また、医用画像データ収集の負担を軽減するため、手術シミュレータ画像を利用した画像基盤モデル構築の方法を確立した。画像ドメイン変換により実画像へと近づけたシミュレータ画像を使用することで、画像基盤モデルの下流タスクにおける性能向上が実現した。そして、開発した画像基盤モデルを使用することで、高精度な手術シーン認識が可能であることを確認した。本研究では、データ収集コストが高い医療分野において、シミュレーションを活用しながら高精度な医療支援 AI を実現するための重要な成果が得られた。

## 1. 共同研究に関する情報

### ⑴ 共同利用・共同研究を実施している拠点名

**名古屋大学 情報基盤センター**

### (2) 課題分野

**大規模計算科学課題分野**

### (3) 参加研究者一覧と役割分担

**小田昌宏（代表）**：医用画像処理コード開発、データ整備

**椋木大地（副代表）**：プログラム最適化、スパコン適用

**片桐孝洋**： ハイパーパラメータ最適化

**森健策**：医用画像処理コード開発

**磯部晃輝**：プログラム最適化、スパコン適用

## 2. 研究の目的と意義

コンピュータによる自動判断を用いた医療支援システムは多く存在し臨床利用されているが、診断の支援を対象としたものばかりであり、治療支援、特に外科手術支援では臨床利用可能な自動支援システムはほとんど実現されていない。外科手術の暗黙知への依存と医師不足の解消を実現するためには、自動的な判断や動作が可能な手術支援ロボットの実現が必須である。

手術支援ロボットの自動化を実現する上で、コンピュータが手術の状況を理解可能とする必要がある。そのために、手術支援ロボットに搭載されたカメラから得られる術中画像を対象として、手術状況理解のための「手術シーン認識」を実現することが求められる。手術シーン認識とは、術中画像から手術の状況を自動認識することを指す。

手術シーン認識を実現する上での難しさは、臓器変化の多様性とデータ不足にある。臓器は柔軟で、術具や周辺臓器の影響を受けて様々な非剛体変形を見せる。複雑な変化を見せる臓器に対応可能な自動認識モデルの実現には、臓器の多様な変形パターンを網羅したデータ収集が必要である。しかし実際の手術からのデータ収集では、データ収集可能

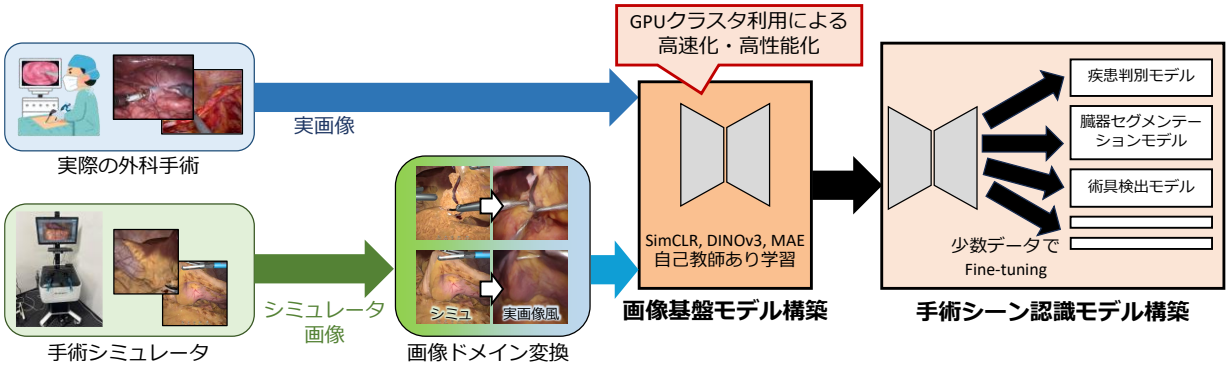



図 1 本研究の概念図。実際の手術画像と手術シミュレータ画像を併用して手術シーンに適した画像基盤モデルを構築する。シミュレータ画像は画像ドメイン変換により実画像へと近づける。画像基盤モデルへの fine-tuning により手術シーン認識モデルを作成する。

な医療機関の数が限られ、さらに手術実施件数が限られていることから、十分なデータを集めるまで数十年と長い時間を要する。そのため手術支援に十分な精度の手術シーン認識は実現されていない。この状況で手術シーン認識手法開発を行うためには、実際の術中画像だけでなく、手術シミュレータで生成したシミュレーション画像を使用するのが有効である。手術シミュレーションを用いることで、様々な臓器変形パターンを含む画像を大量に取得可能となる。シミュレーション画像と実画像の間には見た目（アピアランス）の差異（臓器表面の模様、光沢、術中の焼灼による煙、照明条件の違いなど）があるが、画像ドメイン変換を用いることで軽減可能である。手術シミュレーション画像を用いて手術シーン認識モデルを構築することで、実画像を収集する負担を軽減しながら手術シーン認識モデルを構築可能である。

本研究では、自律的外科手術支援ロボット実現に必須となる、術中画像からの手術シーン認識手法の開発を行った。本研究全体の流れを図 1 に示す。手術シーン認識用深層学習モデルの開発を行う上で、手術シーンに適した画像基盤モデルを独自に開発した。画像基盤モデルを用いることの利点は、(a) fine-tuning において少数のデータを用いるだけで高性能なモデルを実現可能、(b) 様々な手術タスクへ応用可能、の 2 点でありデータ収集コストの高い医用画像処理において有用性が高い。さらに、収集コストの低い手術シミュレーション画像を活用した画像基盤モデル構築方法の確立を行った。そして開発した画像基盤モデルを用いた高性能な手術シーン認識モデル開発を行った。

本研究で得られた重要な成果は下記の 3 点であり、手術支援自動化に大きく貢献するものである。

- **シミュレータ画像併用によるデータ収集の負担を軽減した画像基盤モデル構築方法の確立**
- **手術シーンに適した画像基盤モデル開発**
- **画像基盤モデルを利用した高性能な手術シーン認識モデル開発**

## 3. 当拠点の公募型共同研究として実施した意義

本研究では、高性能な手術シーン認識モデルを実現するため、手術シーン用の画像基盤モデル開発を行った。既存の基盤モデルではなく、対象ドメインに適した画像基盤モデルを独自に開発して利用することで、手術シーン認識の精度向上が実現した。画像基盤モデル構築では、大量の画像データを使用し、膨大な計算量の事前学習処理が必要である。我々が持つ計算資源では、事前学習処理に数か月程度の非常に長い時間を要するため、研究推進が困難であ

った。

本研究を公募型共同研究として実施することで、画像基盤モデル構築における事前学習処理の大幅な高速化が可能となった。スーパーコンピュータが備える GPU クラスタを活用することで事前学習処理が大幅に高速化され、画像基盤モデル構築と手術シーン認識への適用までの成果を得ることができた。

医用画像処理分野では、独自の画像基盤モデル構築と応用を行った例は非常に少なく、本研究の成果は医用画像処理の先駆的な研究成果といえる。この成果は、スーパーコンピュータ利用の有用性を当該研究分野に示す重要なものと考える。

## 4. 前年度までに得られた研究成果の概要

該当なし

## 5. 今年度の研究成果の詳細

本研究のアプリケーションは、術中画像からの手術シーン認識モデルの開発である。近年、画像基盤モデルをベースとして高性能な AI 構築が行われている。画像基盤モデルを利用した AI 開発では、まず大量の画像データを用いて事前学習により画像基盤モデル構築を行い、その後、目的のタスク（下流タスク）用の画像データを使用したモデルの fine-tuning により AI モデルを得る。一度画像基盤モデルが構築されれば、下流タスクにおいては比較的少数の画像データで高性能な AI モデルの構築が可能となるため、データ収集の負担が大きい医療支援において画像基盤モデルを用いる利点は大きい。自然画像向けに様々な画像基盤モデルが提案されているが、手術シーン認識に適した画像基盤モデルはわずかしか提案されておらず、その性能は十分ではない。そのため本研究では、手術シーン認識に適した画像基盤モデルの構築を通して、高性能な手術シーン認識モデルを実現した。画像基盤モデルは他のタスクにも応用可能であり、本研究の成果は様々な手術支援タスクへの発展性が高い。

画像基盤モデル構築の難しさは、大量の画像データが必要である点と、事前学習に膨大な時間を要する点である。これらに対処するため、手術シミュレータと画像ドメイン変換を併用した画像データ収集、並列処理を活用した事前学習の高速化を行った。

研究目的を達成するため、本研究では、**(1) 手術シミュレータの選定**、**(2) 画像ドメイン変換を用いたシミュレータ画像の実画像化**、**(3) 画像基盤モデルの事前学習法の選定**、**(4) 手術シーン用画像基盤モデル構築と手術シーン認識における評価**、を行った。それぞれについて以下で説明する。

### (1) 手術シミュレータの選定

画像基盤モデル構築には実際の術中画像を使用することが望ましいが、大量の実画像収集には時間を要する。そこで、手術シーンを模したシミュレータ画像を画像基盤モデル構築に利用する。

使用する手術シミュレータは、臓器変形の再現度、実画像とのアピアランス（見た目）の類似度が高いものを使用することが望ましい。そこで、既存の研究や市販システムを調査し、外科医による意見も考慮してシステム選定を行った。その結果、他のシミュレータと比較して Simbionix 社の LAP Mentor III が適しているため、本研究で使用した。本機種を用いて腹腔鏡下胃切除術のシミュレーションを行い画像収集を行った。

LAP Mentor III は名古屋大学メディカル xR センターに設置されたものを使用した。本シミュレータで再現可能な術式のうち、腹腔鏡下胃切除を対象とし、医師及び学生がシミュレータを用いて手術を行った際の動画を 30 本収集した。LAP Mentor III の外観とシミュレータから得られた画像の例を図 2 に示す。

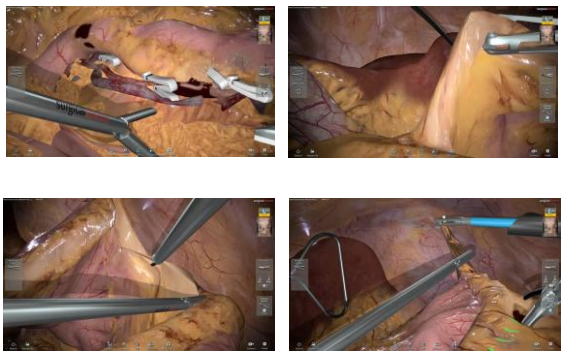
**(2) 画像ドメイン変換を用いたシミュレータ画像の実画像化**

手術シミュレータから得た画像は、臓器表面などのアピアランスが実際の術中画像と異なる。シミュレータと実画像間の差異をなるべく小さくし、シミュレータ画像を実際の術中画像に近づけるため、画像ドメイン変換による画像変換を行った。

画像ドメイン変換手法として、対応のない画像データセット間の画像変換を構築可能な SynDiff (Muzaffer Özbey, et al. IEEE MI 2023) を使用した。画像ドメイン変換では従来 GAN を用いた手法が多かったが、その表現能力には限界があり、学習が不安定という問題があった。SynDiff は、近年の画像生成に多く用いられる Diffusion Model を使用しており、表現力の高さと学習の安定性において優れている。SynDiff の画像変換モデルの学習には、手術シミュレータ画像と実手術画像それぞれ 10,134 枚を使用した。学習処理には 1 ノード 4 GPU (NVIDIA RTX A6000 (48 GB メモリ)) をデータ並列で使用し、学習で約 2 週間を要した。

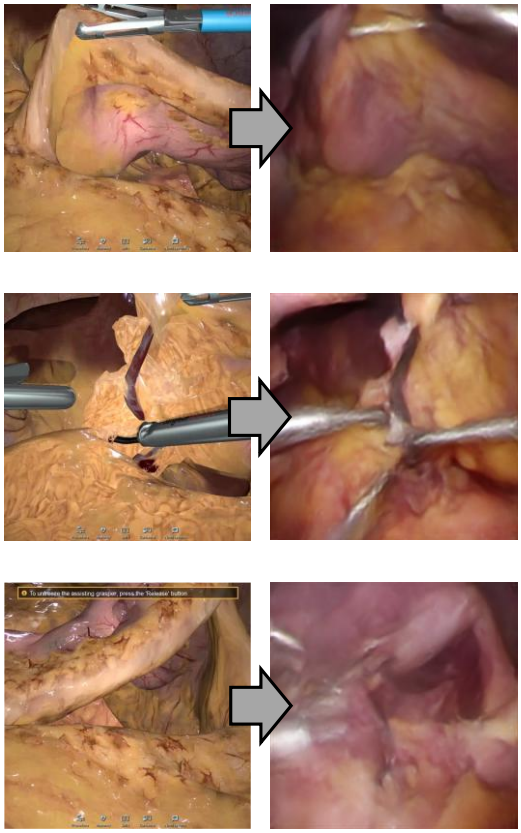
学習後の画像変換モデルを用いて、手術シミュレータ画像 91,206 枚を実画像に近づける変換を行った。その結果を図 3 に示す。変換により、シミュレータ画像の臓器や術具のテクスチャ・光沢を実画像に近づけることができた。この変換後の画像を「**(4) 手術シーン用画像基盤モデル構築と手術シーン認識**


(a)



(b)

図 2 (a) 腹腔鏡下手術シミュレータ LAP Mentor III の外観、(b) 腹腔鏡下胃切除術のシミュレーション画像例。



シミュレータ画像　　実画像への画像ドメイン変換後の画像

図 3 画像ドメイン変換によって手術シミュレータ画像を実画像へ近づける変換を行った結果。

における評価」において使用する。

### (3) 画像基盤モデルの事前学習法の選定

画像基盤モデルの構築では、多くの画像データから自己教師あり学習によりモデルの事前学習を行う。自己教師あり事前学習法として、対照学習ベースの手法（CL）、Masked Image Modeling ベースの手法（MIM）、対照学習と MIM を組み合わせた手法（CL+MIM）が提案されている。これらの手法は数十万枚から数百万枚の画像データセットを用いて事前学習すると良好な画像特徴抽出能力を獲得するが、単一 GPU では事前学習には膨大な計算時間を要する。そこで、GPU クラスタ環境でどれほどの処理の高速化が可能かを評価した。また、医用画像を用いる場合の下流タスクでの性能評価も行った。

#### (3-1) 事前学習手法

使用する事前学習手法について説明する。CL 手法として SimCLR（Ting Chen, et al., ICML 2020）と DINOv3（Oriane Siméoni, et al., arXiv:2508.10104, 2025）を使用した。SimCLR は、画像に対して Data Augmentation による変化を加えながら、変化に頑健な画像表現を獲得する手法である。MIM ベースの手法として Masked Autoencoder（MAE）（Kaiming He, et al., CVPR 2022）を使用した。MAE は画像の一部をランダムにマスクし、エンコーダでマスクされていない領域の特徴抽出、デコーダでマスク部分を復元するタスクを通じて、画像の内容に関する特徴表現を獲得する。CL＋MIM 手法として Contrastive Masked Autoencoder（CMAE）（Zhicheng Huang, et al. IEEE PAMI, 2024）を使用した。CMAE は MAE と対照学習を組み合わせ、画像パッチ間と画像パッチ内の表現学習を行う手法である。

#### (3-2) 事前学習の高速化

まず GPU クラスタ環境における事前学習処理高速化の評価を行った。ここでは事前学習に比較的時間を要する MAE と CMAE を対象とし、PyTorch のデータ並列処理 Distributed Data Parallel（DDP）を用いてマルチノード・マルチ GPU 並列化によるスケーラビリティを評価した。評価では名古屋大学情報基盤センターのスーパーコンピュータ「不老」Type II サブシステムを使用した。各計算ノードは NVIDIA Tesla V100 SXM2（32 GB メモリ）を 4 基搭載している。ノード間通信にはデュアルレール InfiniBand EDR（100 Gbps×2）を使用した。ここでは術中画像の公開データセット（Endoscapes2023＋Cholec80、画像 218,988 枚）を使用した。評価の結果、MAE では 1 ノード（4 GPU）の事前学習時間 1,087 秒が 8 ノード（32 GPU）で 164 秒に短縮され、6.63 倍の高速化と 82.9%の並列効率を達成した。2 ノードで 97.2%、4 ノードで 93.4%の高い並列効率を維持しており、ノード数増加に伴う通信オーバーヘッドの影響は限定的であった。CMAE では 1 ノード 1,085 秒の事前学習時間が 8 ノードで 150 秒となり、7.23 倍の高速化と 90.4%の並列効率を達成した。ただし、最終損失値はノード数増加に伴いわずかに増加する傾向が見られた（0.245 → 0.291）。この傾向は、より多くのエポック数での訓練を考慮すると，単純に時間効率が良いとは言い切れない可能性を示唆している。

#### (3-2) 下流タスクでの性能評価

次に、事前学習済み医用画像基盤モデルの下流タスクにおける性能を評価した。ここでは SimCLR と MAE を対象とする。下流タスクは医用画像分類とし、事前学習済みモデルに対して fine-tuning を行った後に分類精度評価を行った。評価の結果、SimCLR は fine-tuning の学習に使用する画像数が多い（約 13 万枚）場合に約 92%と高い分類精度となったが、fine-tuning の学習に使用する画像数が少ない場合（約 600 枚）には約 70%と分類精度が低下した。MAE は fine-tuning の学習に使用する画像数が多い（約 13 万枚）場合に約 87%であり、fine-tuning の学習に使用する画

像数が少ない場合（約 600 枚）であっても約 75%と高い分類精度であった。これらの結果から、MAE は多くの画像を用いて事前学習を行っておけば、下流タスクにおいて学習用画像数が少なくても高い性能を持つモデルを作成できることが分かった。医療支援においては、画像データが少なくても性能の高いモデルが構築可能であることは非常に有用である。例えば、希少疾患に対する AI モデル構築に有用である。そのため、SimCLR と比較して MAE の利用が適していると言える。

上記の下流タスク評価と同時期に、Meta 社から新たに CL ベースの事前学習法 DINOv3（Oriane Siméoni, et al., arXiv: 2508.10104, 2025）が発表された。DINOv3 を用いて医用画像による事前学習と下流タスクでの評価を実施したところ、fine-tuning の学習に使用する画像数が少ない場合（約 600 枚）において約 77%を達成し、MAE よりも高い分類精度であった。そのため、本研究では画像基盤モデル構築に DINOv3 を利用する。

### (4) 手術シーン用画像基盤モデル構築と手術シーン認識における評価

手術シーン用の画像基盤モデルの構築と評価を行った。事前学習におけるシミュレータ画像利用の効果を検証するため、実画像のみを使用する場合、実画像とシミュレータ画像の両方を使用する場合の両方で画像基盤モデルモデル構築（事前学習）を行った。その後、手術シーン認識の下流タスクにおいて事前学習済みモデルを fine-tuning し、性能評価を行った。

#### (4-1) 画像基盤モデルの事前学習

事前学習の方法について説明する。事前学習法は DINOv3 を使用し、深層学習モデルは ViT-L である。事前学習で使用するデータセットは、(a) 実画像のみ、(b) 実画像+シミュレータ画像、の 2 通りを用意し、2 通りの事前学習を行った。ここで、実画像として術中画像の公開データセット（Endoscapes2023 + Cholec80、画像 218,988 枚）を使用し、シミュレータ画像としては(2)において実画像ドメインへと変換した手術シミュレータ画像 91,206 枚を用いた。

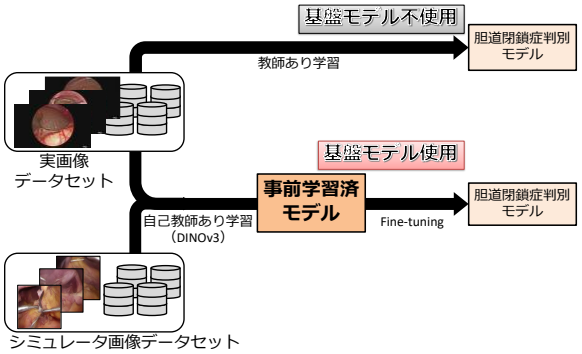



図 4 画像基盤モデルの使用有り・無し 2 パターンでの手術シーン認識モデル構築。

#### (4-2) 画像基盤モデルの下流タスクでの fine-tuning と評価

下流タスクは、肝臓を含む実画像からの胆道閉鎖症症例と正常症例への分類とした。胆道閉鎖症症例では肝臓表面に特徴的な模様（結節状変化）が現れるため、肝臓を撮影した画像から疾患がある程度判別ができる。下流タスクで用いる画像は 1,291 枚であり、これを用いて 2 通りの事前学習済みモデルの fine-tuning と分類精度評価を行った。

分類精度評価の結果、(a) 実画像のみで構築した事前学習済みモデルを使用した場合の分類精度は 89.9%、AUC 0.96 であった。これに対し、(b) 実画像＋シミュレータ画像で構築した事前学習済みモデルを使用すると分類精度は 92.2%、AUC 0.99 となった。実画像に加えてシミュレータ画像を使用することで、下流タスクでの精度向上が確認された。

#### (4-3) 画像基盤モデルを使用しない場合の評価

(4-2)で行った下流タスクでの画像分類モデル構築を、画像基盤モデルを使用せずに行った場合の結果も述べる。モデル構築方法の

違いを図 4 に示す。この場合の分類精度は 79.5%、AUC 0.89 であり、画像基盤モデルを使用する場合と比較して大幅に精度が低くなった。

**(4-4) 評価結果のまとめ**

ここまでの結果より、手術シーン用の画像基盤モデルは高精度な手術シーン認識モデル構築に大きく貢献していた。また、画像基盤モデル構築においては、実画像収集の困難さを軽減するため、大量に生成可能なシミュレータ画像を併用することに効果があることが分かった。この結果は、データ収集が難しい医療支援 AI 開発において、シミュレータ画像の使用によりデータ収集を容易とし、高性能な画像基盤モデルを構築できる可能性を示すものである。

## 6. 進捗状況の自己評価と今後の展望

研究計画に挙げた手術シーン認識モデル実現を目指し、独自の画像基盤モデル開発を通して非常に精度の高い手術シーン認識モデルを実現することができた。最新の事前学習法を導入し、スーパーコンピュータを活用して大規模画像データセットから手術シーンに適した画像基盤モデルを構築した。この画像基盤モデルは、**術中画像を対象としたものの中で非常に高いレベルの性能を実現できた**。また、大規模なデータ収集が困難な医療分野の課題を解決するため、シミュレータ画像を活用した画像基盤モデルの構築方法を確立し、**医療支援 AI の実現を阻む障壁を取り除くことに貢献した**。本研究で得られた知見は、医療支援 AI の性能向上に幅広く貢献するものである。また本研究で得られた成果を国内・国際会議において積極的に発表した。このように、1 年の研究期間で重要な成果を多数挙げることができたため、本研究の達成度は 100%と自己評価する。

(4-2)で得られた画像基盤モデルの評価では、想定通り、シミュレータ画像の併用により性能の高い画像基盤モデルと手術シーン認識モデルが得られた。本研究は、シミュレータ画像利用方法のさらなる工夫による性能向上が期待できる。今回の研究実施で明らかになった課題は、(a) シミュレータ画像の枚数が少なく手術シーン変化の多様性を十分に拡張できていなかったこと、(b) 画像ドメイン変換による変換結果と実画像間のアピアランスの差異がまだ大きかったことが挙げられる。(a) の課題に対しては、様々な手術シーン変化を含むより多くのシミュレータ画像を収集する必要がある。(b) の課題に対しては、画像ドメイン変換モデルの学習においてより多くのデータを使用することと、生成結果に対する条件付きの Diffusion Model を使用し、より実画像らしい結果を得ることが考えられる。このような改善を施すことで、本研究はより優れた手術シーン認識モデルの実現と画像基盤モデル構築が可能になると考える。