

胸腔鏡手術動画内の未知シーンを対象とした U-Net と SAM2 による肺葉セグメンテーションの試み

日高 実咲[†] 門松 由佳^{††} 上野 陽史^{††} 芳川 豊史^{††} 中尾 恵^{†††}

[†] 京都大学医学部人間健康科学科 〒606-8507 京都市左京区聖護院川原町 53

^{††} 名古屋大学医学部附属病院呼吸器外科 〒466-8560 名古屋市昭和区鶴舞町 65 番地

^{†††} 京都大学大学院医学研究科人間健康科学系専攻 〒606-8507 京都市左京区聖護院川原町 53

E-mail: [†]hidaka.misaki.52@st.kyoto-u.ac.jp

あらまし 手術動画内の臓器は画像特徴が類似しており、同一症例でも視野が異なると誤認識が生じやすい。本研究ではこの課題に対し、U-Net と SAM2 を組み合わせたモデルを提案する。U2-Net の予測結果を基に算出したプロンプトを SAM2 に入力することで、未知シーンを対象に肺葉単位の解剖学的構造の自動認識を実現した。さらに、SAM2 の予測結果を用いたプロンプトの更新が認識精度に与える影響についても検討した。

キーワード 術中解剖理解, 協調型セグメンテーション, プロンプト自動生成

Towards Lung Lobe Segmentation for Unseen Thoracoscopic Surgery Videos Using U-Net and SAM2

Misaki HIDAKA[†], Yuka KADOMATHU^{††}, Harushi UENO^{††}, Toyofumi YOSHIKAWA^{††}, and
Megumi NAKAO^{†††}

[†] Human Health Sciences, Faculty of Medicine, Kyoto University, Sakyo, Kyoto, 606-8507, Japan

^{††} Nagoya University Hospital, Syowa, Nagoya, 466-8560, Japan

^{†††} Human Health Sciences, Graduate School of Medicine, Kyoto University, Sakyo, Kyoto, 606-8507, Japan

E-mail: [†]hidaka.misaki.52@st.kyoto-u.ac.jp

Abstract Organs in surgical videos exhibit similar visual features, leading to misrecognition when the field of view changes, even within the same case. To address this issue, we propose a model that combines U-Net and SAM2. By generating prompts from U2-Net predictions and feeding them into SAM2, it achieved automatic lobe-level recognition in unseen scenes. We also investigate how prompt updating based on SAM2 predictions affects recognition accuracy.

Key words Intraoperative anatomical understanding, cooperative segmentation, autonomous prompt generation

1. はじめに

臓器の解剖学的構造の正確な理解は、治療や手術の遂行に不可欠である。外科手術では、血管や腫瘍等の内部構造の位置を把握することにより切除範囲が明確化され、患者及び医師の負担軽減につながる。臨床においては、術前に撮像した三次元 (3D) 医用画像の投影像を手術映像に対して位置合わせをする取り組み [1] が行われてきた。しかし、治療時は手術視野の制限により臓器の全体像を把握することが難しく、特に胸腔鏡手術においては、肺の虚脱による臓器の変形のため、内部の脈管構造や腫瘍位置の正確な同定は一層困難である。

従来の手術シーンに対する認識技術としては、術具の自動



図 1: 胸腔鏡画像の例, (a) 全肺葉と葉間が視認可能なシーン, (b) 中央: 中葉と下葉, 右: 下葉のみが視認可能なシーン

認識 [2] や臓器の輪郭検出 [3] のほか、術前情報から得られる臓器の三次元像を手術シーンに対して正確に位置合わせすることで術中のナビゲーションを可能にする手法 [4] 等がある。この位置合わせの手法においては、手術シーンから臓器領域

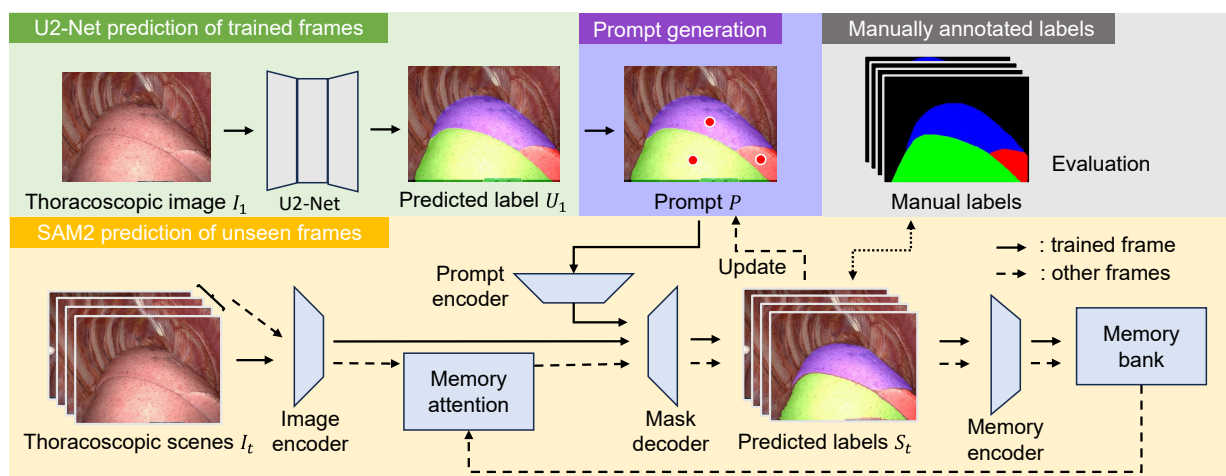


図 2: 提案する U2Net+SAM2 による肺葉単位の領域自動抽出の流れ

を安定して抽出することが重要であり、U-Net [5] を用いたセマンティックセグメンテーションは、位置合わせ精度に寄与する要素となっている [6]。しかし、手術シーンは対象のごく一部の領域が観測された局所情報であり、解剖学的には異なる構造であっても画像特徴が類似している場合が多いため、同一症例内でも視野が変わると誤認識が生じやすい。以下、視認可能な解剖学的単位の数や組み合わせが学習時に用いられたパターンと異なるシーンのことを未知シーンと呼称する。U-Net をあらゆるカメラ移動や臓器の変形等を含む大量のアノテーション付きデータで学習させることで未知シーンに対する臓器領域の認識精度が向上する可能性がある。しかし、臓器の外観や視野の変化は多様であり、手動によるアノテーションは労力を要するため、教師データの集積は容易ではない。このような背景から、手術シーンでの臓器領域の認識において汎化能力を向上させるためには、新たなモデルや工夫が必要である。

近年、動画内の認識対象の領域抽出において汎化能力の獲得を目指したモデルとして、Segment Anything Model 2 (SAM2) [7] が提案され、注目を集めている。SAM2 ではユーザーによる数点の指定により、初見の認識対象に対しても対話的な領域抽出が可能になる。手術シーンでの応用例としては、動画内の術具を対象とした領域抽出の精度改善 [8] が報告されているが、肺葉のように各葉の画像特徴が類似している臓器を対象としたセグメンテーションタスクにおいては、未だ十分な性能が明らかになっていない。また、SAM2 を単独で用いた場合、最初に医師によるプロンプトの入力が必要となるため、完全自動で臓器の領域抽出を実現するには課題が残されている。

本研究の目的は、胸腔鏡手術動画内の未知シーンを対象に肺葉単位の解剖学的構造の自動認識を達成することである。上述の通り手術シーンにおける肺葉単位のセグメンテーションについては、十分な性能が得られづらい。従来手法である U-Net は、単一フレームの画像特徴に基づいて予測を行うため、学習済みシーンでは肺葉単位の自動認識が可能であるが、未知シーンへの対応は困難である。一方 SAM2 は、フレーム間の時系列的連続性に基づいて予測を行うため、未知シーンにおける肺葉単位の領域抽出に対応可能であるが、予測開始時のフレーム

に対してプロンプトの入力を必要とする。本研究では、U-Net の予測結果を基に算出したプロンプトを SAM2 に入力することで、未知シーンに対して肺葉単位の解剖学的構造の自動認識を達成するモデルを提案する。また、SAM2 の予測結果を基にプロンプトを再算出・更新する手法についても検討する。

2. 提案手法

2.1 問題設定

本研究では、京都大学及び名古屋大学の呼吸器外科で胸腔鏡下右肺がん切除術を受けた患者 51 名から取得された、数分間の手術動画を対象とする。図 1 に、動画から取得した胸腔鏡画像の例を示す。提案モデルは、術具を用いた手術操作を開始する前段階における肺葉の解剖学的構造の同定を目的とする。図 1(a) のように全肺葉と葉間が視認可能な胸腔鏡画像を初期シーンと呼称し、初期フレームとしてモデルに与えて予測を開始する。さらに視認範囲が変動する以降の時系列フレームに対しても、肺葉単位の予測ラベルを連続的に取得できる枠組みとして設計する。なお、本研究では、初期フレームに対する学習に U2-Net [9] を使用する。

次に、図 1(b) は視認可能な肺領域の数や葉間の構造が初期シーンと異なる胸腔鏡画像であり、本論文では未知シーンと呼称している。初期シーンに対して十分に学習した U2-Net によっても、未知シーンに対して肺葉単位の十分な認識精度を得ることは難しい。この問題に対し、提案モデルは U2-Net と SAM2 の連携によって自動セグメンテーションを達成しながら、未知シーンに対する認識精度の向上を目指す。

2.2 提案モデルの枠組み

図 2 に、提案モデル (以下、U2-Net+SAM2) の構成及び処理の流れを示す。胸腔鏡動画から取得した時刻 t のフレームを I_t とし、その集合を $I = \{I_t\}$ ($t = 1, \dots, T$) とする。

最初に、初期フレーム I_1 に対して学習済み U2-Net を適用し、肺葉単位の予測ラベル U_1 を得る。得られた U_1 を基に SAM2 に入力するためのプロンプト P を算出する。U2-Net+SAM2 では U_1 に基づいて P を一意に算出することで、未知シーンに対する自動的なセグメンテーションを達成する。プロンプ

表 1: U2-Net と U2-Net+SAM2 における肺葉の認識精度の比較

	mDSC		DSC		
	Overall	Upper	Middle	Lower	
U2-Net	0.8131 ± 0.1109	0.8713 ± 0.0991	0.7641 ± 0.1834	0.8687 ± 0.0835	
U2-Net + SAM2	0.8644 ± 0.0808	0.7909 ± 0.1820	0.8365 ± 0.2015	0.9202 ± 0.0719	

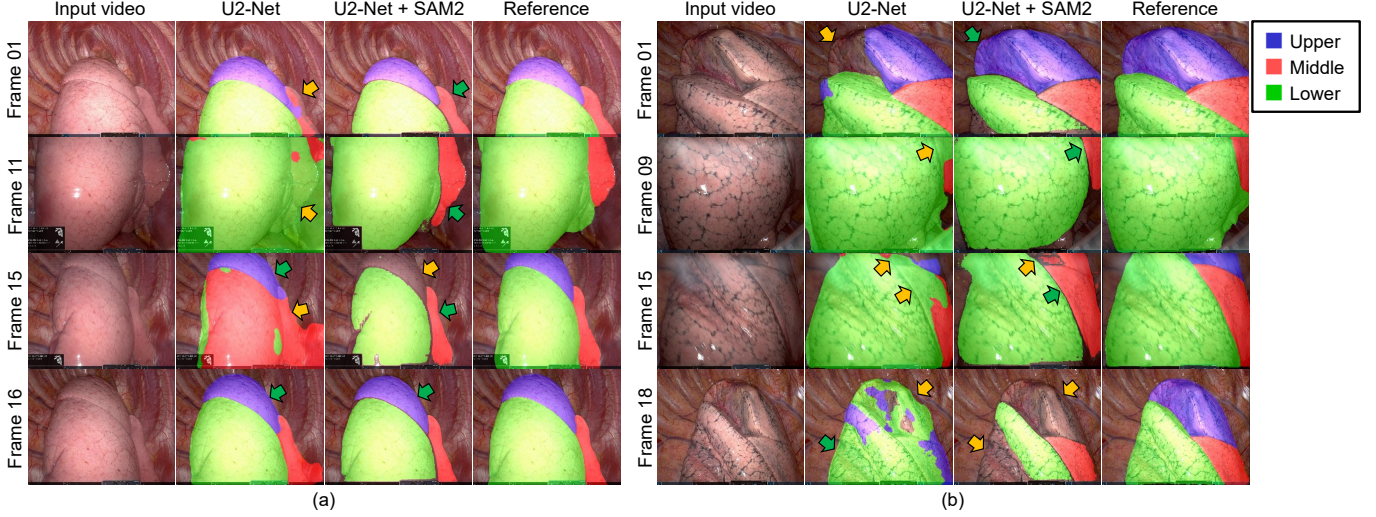


図 3: U2-Net と U2-Net+SAM2 による認識結果の比較（緑矢印：成功，黄矢印：失敗），(a) Case41, (b) Case55

トの自動算出及び更新方法は次節で詳しく述べる。

次に、 I_1 に対して、画像特徴と P の埋め込みを行い、SAM2 により肺葉単位の予測ラベル S_1 を得る。ここで得られた予測情報は内部メモリに保持される。 $t \geq 2$ では、 I_t から得られる画像特徴と過去フレームの予測情報を統合し、 I_t に対する SAM2 による予測ラベル S_t を得る。以上の処理により、予測ラベルの集合として、 $S = \{S_t\} (t = 1, \dots, T)$ を取得する。

2.3 プロンプトの自動算出及び更新方法

本節では、U2-Net+SAM2 におけるプロンプトの自動算出及び更新方法について説明する。図 2 中の P は、 U_1 から自動で算出した各肺葉領域の重心を示しており、対応する肺葉ラベルと合わせて SAM2 に与えるプロンプトとする。

SAM2 による予測では、処理対象のフレーム数が多い場合やシーンの変化が大きい場合に認識精度が低下する可能性がある。この課題に対して、プロンプトを初期フレーム I_1 のみから得るモデルに加え、それ以降のフレームからも取得するプロンプト更新モデルを考える。ここで、未知シーンを含む $t \geq 2$ のフレームでは U2-Net による予測精度は期待できないため、 S_t を用いてプロンプトを更新する。具体的には、 I を一定数の連続フレームごとに分割し、各区間を chunk と定義する。各 chunk 内では固定されたプロンプトを用いて SAM2 による予測を行い、chunk の末尾フレームに対する各肺葉の予測ラベルから重心を得る。得られたラベル付き重心群を次の chunk に対するプロンプトとして SAM2 へ与える。

3. 実 験

3.1 対象データと実験条件

本研究では、2.1 節で導入した手術動画 51 症例のうち、40

例を初期シーンに対する U2-Net の学習と検証に、11 例を未知シーンに対する U2-Net+SAM2 の検証に使用した。

U2-Net の学習及び検証のために、40 症例の各手術動画から全肺葉と葉間が視認可能なフレームを 3-27 枚抽出し、取得した計 531 枚のうち 479 枚を学習用、52 枚を検証用に使用した。各フレームに対してトリミングとパディングによって縦横のサイズを揃えた後、 256×256 pixel にリサイズした画像を U2-Net に入力した。U2-Net の学習では、主出力及び中間レイヤの出力に対して算出した Cross Entropy の総和に、重み 0.01 の正則化項を加えた損失関数を用いた。バッチサイズは 16、エポック数は 30 に設定し、同一条件下で学習を 10 回繰り返した。各学習において検証用データに対する認識精度を評価した上で、最も高い精度を示したモデルの重みを最終モデルとして採用した。評価指標には、手動で作成した肺葉単位のラベルと U2-Net の予測ラベル間で算出される mIoU (mean Intersection of Union), mDSC (mean Dice Similarity Coefficient) を用いた。検証の結果、mIoU: 0.9482, mDSC: 0.9728 の評価値が得られ、初期フレームにおいては U2-Net による高精度な肺葉単位のセグメンテーションが可能であることが確認された。

U2-Net+SAM2 の検証のために、未知シーンを含む 11 症例の各手術動画から 8-41 枚のフレームを抽出し、計 233 枚の画像を取得した。本研究では、術具による手術操作が開始される前の肺葉認識を目的とするため、術具が肺表面の一部を遮蔽している画像は評価対象から除外した。この除外処理により、29.97 FPS (Frame Per Second) の各動画内のフレームから 4-45 の不均一な間隔 (約 0.13-1.5 秒に相当) で抽出されたフレーム群を評価対象とした。また、SAM2 は画像下部の情報提示枠を独立した物体として認識するため、下部 25 %につ

表 2: プロンプトの入力条件の違いによる認識精度の比較

	mDSC	DSC		
	Overall	Upper	Middle	Lower
U2-Net + SAM2	0.8644 ± 0.0808	0.7909 ± 0.1820	0.8365 ± 0.2015	0.9202 ± 0.0719
U2-Net + SAM2 (c=1)	0.7688 ± 0.2281	0.8028 ± 0.1909	0.7743 ± 0.2782	0.7984 ± 0.2837
U2-Net + SAM2 (c=3)	0.8633 ± 0.0812	0.7970 ± 0.1857	0.8316 ± 0.2064	0.9248 ± 0.0865
U2-Net + SAM2 (c=5)	0.8578 ± 0.0802	0.7965 ± 0.1857	0.8277 ± 0.2012	0.9194 ± 0.0878

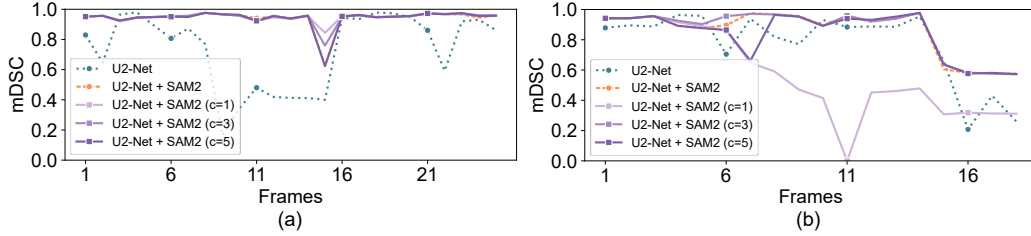


図 4: プロンプトの入力条件の違いが時系列フレームの認識精度に与える影響, (a) Case41, (b) Case55

いても評価対象外とした。なお、本実験においては、SAM2 は公開されている学習済みのモデルをそのまま使用し、評価指標には mDSC, 各葉の DSC, 予測時の FPS を用いた。

3.2 U2-Net + SAM2 の性能評価

最初に、初期フレームに対してプロンプトを 1 回のみ算出する条件で、U2-Net+SAM2 による肺葉単位の認識精度を検証した。表 1 に、U2-Net と U2-Net+SAM2 による mDSC, 各肺葉の DSC の平均値を示す。U2-Net+SAM2 は U2-Net と比較して、mDSC の向上が確認された。各葉の DSC に着目すると、中葉及び下葉では上昇し、上葉では低下した。また、U2-Net では 35.0 FPS であるのに対し、U2-Net+SAM2 では 2.36 FPS と大きく低下した。

図 3 に両モデルによる肺葉単位のセグメンテーション結果を示す。U2-Net+SAM2 は U2-Net と比較して、mDSC が Case41 では約 28.4 %, Case55 では約 10.9 % 上昇した。肺葉ごとに着目すると、カメラ視野の移動、さらに肺表面の変形や変色の影響により、認識が不安定となる傾向が確認された。

3.3 プロンプト更新の有無が認識精度に与える影響

次に、U2-Net+SAM2 においてプロンプトの更新が肺葉単位の認識精度に与える影響を調査した。ここで、chunk の長さを c と表し、 $c=1,3,5$ に設定して実験を行った。表 2 に、各条件における mDSC, 各葉の DSC の平均値を示す。プロンプト更新による認識精度の大きな改善は確認されなかったが、一部の条件では、肺葉単位の認識精度が僅かに改善した。また、更新間隔が短いほど FPS が低下する傾向が見られた。

図 4 に各条件におけるフレーム単位の mDSC を示す。症例によっては、プロンプトの更新により一部のフレームで mDSC の低下が抑制されたが、更新頻度が高い場合には、低精度の予測に基づく更新により誤差が蓄積される可能性が示唆された。

4. おわりに

本研究では胸腔鏡手術動画の未知シーンに対し、U2-Net と SAM2 を組み合わせた肺葉単位の自動認識の手法を提案した。

実験の結果、提案モデルは U2-Net と比べて高い認識精度を示すことを確認した。さらに、SAM2 の予測結果を用いたプロンプトの更新により一部の症例において精度が改善することが示唆された。

文 献

- [1] Y. Kadomatsu, M. Nakao, H. Ueno, et al. Clinical application of resection process map as a novel surgical guide in thoracic surgery. *Interdisciplinary CardioVascular and Thoracic Surgery*, Vol. 36, No. 4, p. ivad059, 2023.
- [2] N. N. Myo, A. Boonkong, K. Khampitak, et al. Real-time surgical instrument segmentation analysis using YOLOv8 with Bytetrack for laparoscopic surgery. *IEEE Access*, Vol. 12, pp. 83091–83103, 2024.
- [3] B. Koo, M. R. Robu, M. Allam, et al. Automatic, global registration in laparoscopic liver surgery. *International Journal of Computer Assisted Radiology and Surgery*, Vol. 17, No. 2, pp. 167–176, 2022.
- [4] N. Haouchine, P. Juvekar, M. Nercessian, et al. Pose estimation and non-rigid registration for augmented reality during neurosurgery. *IEEE Transactions on Biomedical Engineering*, Vol. 69, No. 4, pp. 1310–1317, 2022.
- [5] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Proc. Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pp. 234–241, 2015.
- [6] M. Kobayashi, Y. Kidoguchi, S. Ogiso, et al. Shape registration for laparoscopic images using offline diffusion learning. In *International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2025.
- [7] N. Ravi, V. Gabeur, Y.-T. Hu, et al. SAM 2: Segment anything in images and videos. In *Proc. International Conference on Learning Representations (ICLR)*, 2024.
- [8] H. Liu, E. Zhang, J. Wu, et al. Surgical SAM 2: Real-time segment anything in surgical video by efficient frame pruning. In *Proc. Conference on Neural Information Processing Systems (NeurIPS)*, 2024.
- [9] X. Qin, Z. Zhang, C. Huang, et al. U²-Net: Going deeper with nested u-structure for salient object detection. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10766–10775, 2020.