# ロボット支援手術動画からの転移学習による顕微鏡縫合の熟練度推定と時系列可視化

佐藤 響[1,a)]　清藤 哲史[2]　宮藤 詩緒[1,b)]

**概要**：本研究は，顕微鏡縫合の学習者に自身の能力を客観的に把握できるフィードバックを与える学習支援を目的として，少ない訓練動画からでも顕微鏡縫合の熟練度（初心者/中級者/上級者）を機械学習モデルにより高精度に評価する枠組みを提案する．医療分野では特定タスクの大規模な学習データを集めることが困難という課題に対し，公開されているロボット支援手術の動画データセット「JIGSAWS」で事前学習したモデルを顕微鏡縫合タスクに転移させることで対応した．動画を複数のクリップに分割してスコアを算出し，アイソトニック回帰で補正後に 3 クラスの熟練度に分類を行う．検証の結果，本手法は顕微鏡縫合データを用いてゼロから学習させた場合や，一般的な動作動画データで事前学習済みのモデルから顕微鏡縫合データで学習させた場合と比較して最も高い精度（Accuracy=0.750，QWK=0.679）を達成した．これにより，関連分野のデータを用いた転移学習が，少数のデータセットにおける技能評価の精度向上に有効であることが示された．また，技能レベルを時系列で可視化するヒートマップも作成したが，医師のレビューにより，非動作区間の扱いや微細な動きの認識など信頼性に改善の余地があることも明らかになった．

## 1. はじめに

脳神経外科医の技能訓練，とくに顕微鏡縫合は高い精度と安定性を要求するため数万針という相当な訓練が必要である [1]．一方，学習者が受け取る評価は指導者の主観性に左右されやすく，指導者への負担も大きい．これに対し，個人訓練を支援するための各種訓練支援手法が提案されてきた [2–6]．なかでも，スコアリングシステムや習熟度推定による訓練評価は，従来学習者個人の主観的評価しかなかった個人訓練に対し，客観的な評価基準を導入できる点で有用である [4,5,7]．しかし，医療現場ではその専門性の高さから，大規模にラベル付けされた動画データの収集が容易ではない．また，本研究では専門家による評価の代替として被験者の経験年数に基づきラベルを付与するため，技能指標としては粗く，信頼性の限界も伴う．このため，限られたデータ条件でも安定に機能しうる熟練度推定手法が求められる．

本研究では，顕微鏡縫合の学習者が「どこが良く／悪かったか」を自己確認できるフィードバック提示（時系列可視化）を主目的として，動画から熟練度を推定し，動画内で可視化を行う訓練支援手法を提案する．本研究の中心的な問いは，「少数の顕微鏡縫合動画しか得られない状況で，学習者へのフィードバックに耐える一貫性を持ったスコア列（時系列）と，それを要約した動画単位の熟練度を，どこまで信頼して提示できるか」である．

少数データで推論を行うための具体的な手法として，まず，ロボット支援手術データセット JIGSAWS（Suturing/Knot-Tying）で動画表現を事前学習し，顕微鏡縫合タスクへと転移学習を行う．推論は動画から等間隔に取得した 8 クリップの出力を平均して動画スコアを得て，2 つの閾値で 3 クラス分類に離散化する．さらに，スコアの時系列ヒートマップを作成し，熟練医師によるレビューで可視化の妥当性と教育的有用性を調べる．検証は被験者独立（LOUO）で行い，Accuracy と QWK を主指標とする．比較は，(i) 提案（JIGSAWS 事前学習＋校正），(ii) 顕微鏡縫合のみで学習（ランダム初期化），(iii) 顕微鏡縫合のみで学習（Kinetics-400 初期化）の 3 条件を評価する．

本研究の貢献は以下の通りである．

1. **少数データ設定での熟練度分類**：学習支援インタフェース設計のための知見として，JIGSAWS 由来の事前学習と校正の組合せにより，少数の顕微鏡縫合データでも動画単位の分類精度を向上できることを示す（LOUO にて Accuracy=0.750，QWK=0.679）．

[1] 東京科学大学情報理工学院
[2] 富士脳障害研究所附属病院脳神経外科
[a)] sato.h.7b51@m.isct.ac.jp
[b)] miyafuji@comp.isct.ac.jp

2. **時系列スコアの可視化と実務検証**：ヒートマップと医師レビューにより，可視化が示す要改善区間の妥当性と，運用上の課題（非動作区間の扱い・微細運動の捉えにくさ）を特定する．
3. **転移学習の有効性の検証**：ロボット支援手術（内視鏡視点）から顕微鏡縫合（顕微鏡視点）へのロボットから人間へというドメインギャップを前提に，外部データで獲得した時空間特徴が有利に働くことを定量的に示す（Kinetics-400 初期化およびランダム初期化を上回る）．

## 2. 関連研究

### 2.1 医療データセットの活用と技能評価

外科技能の客観評価は，従来の OSATS [8] や GRS [9] などの専門家評価指標を基盤として発展してきたが，AI の発展により自動化・大規模化が進んでいる．とくに動画ベース評価（Video-Based Assessment; VBA）は，腹腔鏡・内視鏡・顕微鏡・ロボット支援手術など，カメラを用いる広範な術式で入手容易かつ技能判定に必要な情報を多く含むことから，有効性と実運用可能性の両面で注目されている [10]．この潮流を牽引してきた代表的データセットが，JIGSAWS（JHU-ISI Gesture and Skill Assessment Working Set）[11] である．JIGSAWS はロボット支援下の訓練課題（Knot-Tying/Suturing/Needle-Passing）に対してビデオと運動学の両方を収録し，専門家スコアによる注釈を備える公開スキル評価データセットとして広く用いられてきた．このデータセットを用い，運動学の時系列をそのまま DL に入力して熟練度 3 分類を行う枠組みが提案され，Suturing/Needle-Passing/Knot-Tying で 92.5%/95.4%/91.3%といった高精度が報告されている [12]．

JIGSAWS を動画（2D）側から扱った研究では，光フロー由来の特徴や 3D-CNN/LSTM 等を用いた動画ベース技能評価（VBSA）が体系的に検討され，JIGSAWS の 3 課題すべてで 80%超の平均精度などの報告がある [13]．ただし，ドメイン差による性能崩れ（例：別データで学習した器具分割モデルが JIGSAWS では過学習的挙動を示す）も指摘され，データセット間の差異が汎化のボトルネックになりうるといわれている．

### 2.2 少数データ学習と転移学習

医療動画は，取得・匿名化・注釈の負担が大きく，ラベル付きデータが少量に留まりがちである．深層学習をそのまま適用すると過学習に陥りやすいため，まずは 転移学習（事前学習→微調整） が基盤戦略として確立してきた [14,15]．一般動画で訓練された 3D CNN（例：R(2+1)D-18）や，動作データセット（例：Kinetics-400）の事前学習重みを初期化に用いることで，少数データでも収束を安定化させ，適度な正則化効果を得られることが広く報告されている [16,17]．もっとも，ソースとターゲットのドメイン関連性が低いと転移効果は限定的になる [14]．この関連性の問題に対し，近年は自己教師あり学習（SSL）による動画事前学習が注目されている．VideoMAE のようなマスクド復元ベースの手法は，ラベル不要で運動・形状の普遍表現を学習でき，少量ラベルの下流タスクでも堅調に働くことが示されている [18]．その他にも例えば，転移学習に頼らず，タスクに合わせたよりシンプルなモデルをゼロから学習させる方法がある．また，ピクセル単位の正確なラベルの代わりに，画像全体に対する診断ラベル（例：症例の有無）といった，より入手しやすい弱いラベルを活用する多インスタンス学習（MIL）などの手法も有効な選択肢となる [14,15]．外科領域でも，EndoNet 系の研究が示すように，術具・位相（フェーズ）といったドメイン固有の視覚要素を取り込んだ表現が，タスク特異の識別に寄与する．

さらに，外科技能評価においては，対象タスクの少数データで直接学習を行うのではなく，公開データセットを事前学習元として活用し，異なる環境（ドメイン）へ知識を移転させるクロスドメイン転移の試みも報告されている．Zhang ら [19] は，ロボット支援顕微鏡手術（RAMS）の技能評価に対し，JIGSAWS を用いて事前学習したモデルをターゲットドメインへ適応させる枠組みを提案している．同研究では，主に運動学系列を入力としつつ，術具追跡による空間的なフィードバックの提供や，Grad-CAM を用いた判断根拠の可視化を併用している．また，ラベル空間は共通するが入出力分布が異なる状況をトランスダクティブ転移（ドメイン適応）として位置づけている．これに対し本研究は，運動学情報の取得が困難な顕微鏡縫合の訓練動画を対象とする点を前提に，JIGSAWS を動画ベースの事前学習元として活用し，顕微鏡下の手技動画へ転移させる．すなわち，ロボット支援手術（内視鏡視点）から人手による顕微鏡視点へと，観察視点や器具外観が大きく異なるロボットから人間へというドメインギャップを扱う点，および技能推定を時系列の可視化によるフィードバックへと繋げている点に独自性がある．以上のように，本研究は JIGSAWS を用いた転移学習の系譜に位置づけつつも，入力モダリティ（運動学 vs 動画）およびドメインギャップの性質（ロボット間 vs ロボットから人間）において既存研究と明確に異なる．

### 2.3 時系列スコアによる良否区間の同定と学習者フィードバック

本節では，動画から時間方向の技能スコア列を推定し，良い区間/悪い区間を特定して学習者に返す研究の流れを概観する．前提として，AQA（Action Quality Assessment）は体操や飛込などの採点競技を起源とする行為の質の推定課題であり，最終的に一つの総合点を出すだけでなく，時

間軸上のどこが高品質でどこが低品質だったかを把握できることが利用上は重要である．ところが，実際の採点は人間の評価者（審査員）に依存し，評価者間のばらつきやラベルの曖昧さが不可避に存在する．この前提を正面から扱い，単一の正解値に回帰するのではなく，時刻（クリップ）ごとのスコアを確率分布として学習することで，不確実性を明示的に織り込みつつ時間方向の「山/谷」を安定的に可視化するアプローチが示された [20]．同じく「どこが決め手か」を説明する流れで，長尺動画向けに採点の内在ロジック（区間の重要度×出来栄え）をモデル側に分離して推定する Weight–Score（二系統）回帰が提案され，さらに Transformer の注意機構が時間的に飛び飛びになって情報を均してしまう現象（Temporal Skipping）を抑える損失設計が加えられた．これにより「どの区間をどれだけ重視し，その区間を何点と判断したか」を連続時間上に可視化でき，学習者へのフィードバックに直結する粒度で根拠を示せるようになった [21]．同様の課題意識は外科スキル評価の分野にも見られ，JIGSAWS を用いて最終点に加え進行中の中間スコア（running intermediate scores）を推定し，問題のあるジェスチャ区間を局在化したうえで，強化学習により望ましい操作系列（改善方策）を提示する枠組みが示されている．すなわち「いつ悪かったか」だけでなく「なぜ悪かったか/どう直すか」までを一連の出力として返し，教育支援の実装レベルに踏み込んだ [22]．さらに最近では，Video–Language Model を用いた看護手技解析のように，粗い手順同定から細粒度サブアクションの分解，欠落操作や順序誤りの検出へと段階的に推論を進め，自然言語の説明付きフィードバックを生成する枠組みが登場している [23]．

## 3. 提案手法

### 3.1 概要

本研究は，手術技能に関する公開データセットを用いた事前学習と，評価対象である顕微鏡縫合訓練動画に対する校正および時系列解析を統合した熟練度評価パイプラインを提案する．図 1 に全体構成を示す．

本研究では熟練度を 3 段階（初心者/中級者/上級者）の離散値として定義する．これは，評価用データが小規模であり，かつ経験年数に基づく粗いラベルを用いるという制約の下では，連続値回帰やより細かい多クラス分類は学習の不安定化や過学習のリスクが高いと考えられるためである．そのため，本研究では頑健な識別性能を確保しうる最小限の分解能として 3 クラス分類を採用する．これは，同様に小規模な外科手術動画データに対して転移学習による 3 クラス分類を行い，その有効性を検証している先行研究 [19] の構成とも整合するものである．

熟練度評価の具体的な手順としては，まず，公開データセットを用いて時空間特徴抽出器（バックボーン）を事前学習し，次いで顕微鏡縫合データを事前学習モデルに入力して得られるモデル出力に対してアイソトニック回帰に基づく出力校正と閾値最適化を行い，最終的な 3 クラス（初心者/中級者/上級者）予測を得る．推論時には動画全体からスライディングウィンドウを用いてクリップを連続的に取得し，予測結果を並べることでスコア時系列データを生成する．

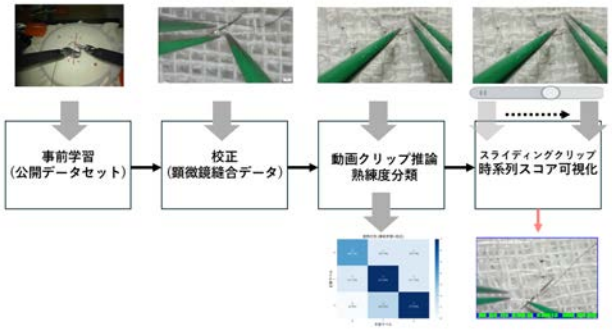



図 1: 提案手法ワークフロー

### 3.2 事前学習ステージ

本研究では，公開手術技能データを用いたマルチタスク事前学習により，クリップ（短時間区間）と動画全体の双方に整合的な時空間表現を獲得する．特徴抽出には，R(2+1)D-18 [16] アーキテクチャを採用し，実装には PyTorch [24] の torchvision ライブラリで提供されているモデルを使用した．入力はサイズ 224 × 224 の動画クリップであり，フレーム数は T=16 とする．元動画（30fps）から 3 フレームごとにサンプリングしてクリップを構成するため，クリップの実効サンプリングレートは 10fps，相当時間長は約 1.5 秒（=(16-1)/10）となる．学習は以下の 3 つのタスクについて行う．

- クリップ単位のジェスチャ分類
- 動画単位のスコア (Global Rating Score:GRS) 回帰
- クリップの貢献度予測（各クリップが動画全体スコアへ与える寄与の推定）

クリップ抽出は，事前学習では各動画から等間隔に 8 クリップを取得し，動画全体を均等にカバーする設定とした．一方，比較条件として顕微鏡縫合のみで学習する場合は，クリップ長・サンプリングレートは同一としつつ，学習時の切り出し位置をランダム化した．また，転移学習へ向けた汎化性能の獲得を目的とし，学習時の前処理にランダム拡大縮小・幾何/色摂動を適用した．

### 3.3 損失関数と学習

損失関数は，転移学習へ向け汎用的な時空間表現の獲得を目指して設計された．

総損失は「ジェスチャ分類（KL）＋ GRS 回帰（MSE）＋

貢献度の整合制約＋正則化（L1, TV)」の和として定義される．ここで $i$ はクリップのインデックス，$N$ は 1 動画から抽出するクリップ数を表す．$p_i$ はクリップ $i$ に対するジェスチャ分類の予測確率分布，$\hat{p}_i$ は正解ラベルに対応する分布である．$\mathrm{KL}\left(p_i \parallel \hat{p}_i\right)$ は両分布の Kullback–Leibler ダイバージェンスであり，$\mathbb{E}_{i\in\mathcal{V}}\left[\cdot\right]$ は学習に用いたクリップ集合（ミニバッチ）$\mathcal{V}$ 上での平均を表す．また，$y$ は動画単位の真の GRS(Global Rating Scale: 熟練度スコア)，$\hat{y}$ はモデルが予測した動画単位 GRS である．貢献度 $c_i$ は，各クリップ $i$ が動画全体のスコアへ与える寄与を表すモデルの出力である．整合制約では「全クリップの寄与の総和が，モデルが予測した動画スコア $\hat{y}$ と整合する」ことを課し，局所評価（どの区間が効いたか）と全体評価（総合スコア）の整合性を保ちながら学習させる．この際，$\hat{y}_{detach}$ を用いることで，整合制約の勾配が逆伝播することを防いでいる．さらに L1 正則化項によって，クリップの貢献度をスパース化 (多くを 0 に近づける) させ，スコアを大きく左右する重要な瞬間を強調する．最後に TV（Total Variation）正則化項によって時間的な変化を平滑化し，一連の動作として自然な貢献度の変化を促している．係数 $\lambda$ は項の相対的な寄与を調整するハイパーパラメータである．本研究ではこれらの重みを，$\lambda_{gest}=1.0$，$\lambda_{grs}=0.5$，$\lambda_{\Sigma}=0.5$，$\lambda_1=0.01$，$\lambda_{TV}=0.1$ と設定した．

$$
\begin{aligned}
\mathcal{L} &= \lambda_{\mathrm{gest}}\mathcal{L}_{\mathrm{gest}} + \lambda_{\mathrm{grs}}\mathcal{L}_{\mathrm{grs}} + \lambda_{\Sigma}\mathcal{L}_{\Sigma} + \lambda_1\mathcal{L}_1 + \lambda_{\mathrm{tv}}\mathcal{L}_{\mathrm{tv}} \\
&= \lambda_{\mathrm{gest}} \cdot \mathbb{E}_{i\in\mathcal{V}}\left[\mathrm{KL}\left(p_i \parallel \hat{p}_i\right)\right] \\
&\quad + \lambda_{\mathrm{grs}} \cdot (y-\hat{y})^2 \\
&\quad + \lambda_{\Sigma} \cdot \left(\sum_{i=1}^{N} c_i - \hat{y}_{\mathrm{detach}}\right)^2 \\
&\quad + \lambda_1 \cdot \frac{1}{N}\sum_{i=1}^{N} |c_i| \\
&\quad + \lambda_{\mathrm{tv}} \cdot \frac{1}{N-1}\sum_{i=2}^{N} |c_i - c_{i-1}|
\end{aligned}
$$

### 3.4 校正（Calibration）と離散化

転移学習先の顕微鏡縫合データに合わせるため，事前学習モデルの連続出力をアイソトニック回帰で単調校正する．校正スコアに対して 2 つの閾値 $t_0, t_1$ を探索し，3 クラスに離散化する．閾値は QWK を最大化するよう決定される．

### 3.5 スコアの時系列可視化

訓練中の「うまくできている部分」と「改善すべき部分」を時間軸上で明示し，学習者にフィードバックすることを目的とする．動画を時間方向にスライディングしながら連続的にクリップを取得し，各時刻のクリップ予測を並べて時系列スコア列を得る．各点は 3.2 節と同一のクリップ設定（T=16，3 フレーム間隔，窓幅約 1.5 秒）で推定される．スライディング間隔 $\Delta$ は元動画 30fps に対して 8 フレーム（約 0.27 秒）とした．このとき隣接クリップは約 82% 重複し，動作の微細な変化を見落としにくい一方で，予測値の一時的な跳ね（スパイク）や閾値付近での頻繁な入れ替わり（チャタリング）が生じやすい．

そこで本研究では，可視化および区間抽出に先立ち，以下の時系列後処理を適用する．

1. **移動平均による平滑化**：前後を含む 3 点のスコアの平均を取り，短周期のノイズを抑えて全体トレンドを抽出する．
2. **判定状態の安定化**：離散化の閾値に対してマージンを設けたヒステリシス判定を用い，閾値近傍の微小変動による頻繁なクラス遷移を抑制する．
3. **最小持続時間フィルタ**：抽出された区間が 0.5 秒未満の場合はノイズとして除去し，人間が解釈可能な時間スケールの区間のみをレポートとして出力する．

以上により，可視化の滑らかさと判定の安定性を両立させる．得られた系列は時間的ヒートマップとして表示し，閾値に基づき良好区間/改善区間をクラス色（赤=初級/緑=中級/青=上級）で可視化する．

## 4. 実験

本研究では，提案する転移学習モデルによる顕微鏡縫合動画の熟練度推定性能を検証するために 2 段階の実験を実施した．第一に，ロボット支援手術動画から事前学習を行ったモデルを用いて，顕微鏡縫合動画に対する熟練度分類を行い，その精度を既存手法と比較評価した．これにより，提案手法が少数データ環境下でも有効に機能するかを検証する．第二に，推定されたスコアを動画中の時系列に沿って可視化し，その可視化結果を一名の熟練医師に提示して自由記述によるレビューを依頼した．これにより，可視化された熟練度推定が実際の医師の観点から見て妥当であるかを評価する．これらの実験を通じて，提案システムの精度的妥当性と専門家評価に基づく実用的妥当性の双方を明らかにすることを目的とする．

### 4.1 データセット

#### 4.1.1 JIGSAWS(事前学習用)

事前学習には，内視鏡カメラで撮影された外科ロボットの手術動画データセット JIGSAWS [11] を用いる．データセットに含まれるタスクは Suturing / Knot-Tying / Needle-Passing の 3 種だが，縫合技能の評価に直接的に関係が深い Suturing と Knot-Tying の 2 種のみを採用する (図 2(a,b))．これらは糸・針の取り回しや結紮操作を含み，顕微鏡下縫合における動作と運動学的に共通点が多い．

撮影は左右のカメラで行われているが，視点を統一する


(a) JIGSAWS(Suturing)


(b) JIGSAWS(Knot Tying)

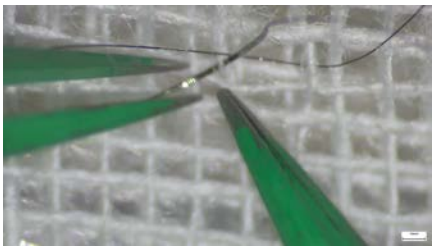

(c) 顕微鏡縫合訓練

図 2: 3 種の手術タスク動画

表 1: 事前学習用データ概要

| 項目 | 内容 |
|---|---|
| 被験者数 | 8 名 |
| 動画数 | 75 本 |
| フレームレート (fps) | 30 |
| ラベル 1（スコア） | 動画ごとの技能スコア (6〜30 点) |
| ラベル 2（ジェスチャ） | 時刻ごとのジェスチャ分類 (15 種類) |

表 2: 評価用データ概要

| 項目 | 内容 |
|---|---|
| 被験者数 | 14 名 |
| 動画数 | 24 本 |
| フレームレート (fps) | 30 |
| ラベル (熟練度) | 初心者：3 年未満<br>中級者：8 年未満<br>上級者：8 年以上 |

ため左カメラ動画のみを使用する．その他の情報を表 1 に示す．

#### 4.1.2 顕微鏡縫合データ (評価用)

評価には，ガーゼ繊維を結ぶことで微小組織の結紮動作を模擬した顕微鏡縫合訓練タスクの動画 (図 2(c)) を用いる．動画は顕微鏡に取り付けられた顕微鏡カメラ（ZEISS Axiocam 208 color）から撮影された．被験者 14 名分の顕微鏡縫合動画 24 本（合計約 120 分）を用意し，熟練度は経験年数に基づき 3 段階（初心者/中級者/上級者）に付与した (表 2)．内訳は初心者 6 本，中級者 9 本，上級者 9 本である．

### 4.2 動画の熟練度評価と精度検証

評価は動画単位で行う．各動画の予測は，均等な間隔で取得した 8 クリップの校正スコアの平均を用い，最適化した 2 つの閾値 $t_0, t_1$ を適用してクラスに分類する．

そして，LOUO（Leave-One-User-Out） でクラス分類の精度評価を行う．具体的には，被験者一名を検証用に保持し，その被験者の全動画を学習から除外する．残り被験者のデータのみで 3.4 節の校正（アイソトニック回帰）と閾値最適化を実施し，学習から除外した被験者の動画で評価を行う．この手順を全被験者について反復し，得られた結果を集計する．

### 4.3 熟練度評価の比較条件

提案手法の有効性を検証するため，以下の比較条件を設定した．いずれも R(2+1)D-18 をバックボーンとし，顕微鏡縫合訓練動画のみで学習する．ヘッドには CORAL（COnditional RAnk Logits）[25] を用いる．CORAL はランクが $r$ より大きいかを判断するための $K-1$ 個のタスクを同時に解くことによりクラス分類を行う手法で，これにより順序性 (初心者＜中級者＜上級者）をモデルが自然に学習する．今回は $r$ より大きいかを判断する閾値は 0.5 で固定とした．これを重みの初期値をランダムにしたものと，より汎用的な動きの認識能力を持つものとして Kinetics-400 [17,26] で事前学習済みの重みを使用したものを 2 つ用意し比較を行う．最適化手法は AdamW，バッチサイズは 2 とし，他の設定は比較条件間で共通とした．ただし初期化方法により学習率と学習エポック数を変更している．ランダム初期化では epoch=100，学習率 LR=3e-4 とした．Kinetics-400 初期化では事前学習済みであることを考慮して epoch=50，LR=3e-5 とし，バックボーンの学習率倍率を 0.1 に設定した．さらに，学習初期の過学習を避けるため，最初の 5 エポックはバックボーンを凍結した．評価は 4.2 節と同様に 8 クリップの結果の平均からクラスを分類し，LOUO にて行った．

比較条件をまとめると以下のようになる．

- **JIGSAWS 事前学習＋校正 (提案)**
  JIGSAWS で事前学習後，顕微鏡縫合動画で校正および閾値決定
- **顕微鏡縫合訓練動画で学習（ランダム初期化）**
  バックボーンを **ランダム初期化**し，顕微鏡縫合訓練動画のみで学習
- **顕微鏡縫合訓練動画で学習（Kinetics-400 で初期化）**
  バックボーンを **Kinetics-400 事前学習済み重み**で初

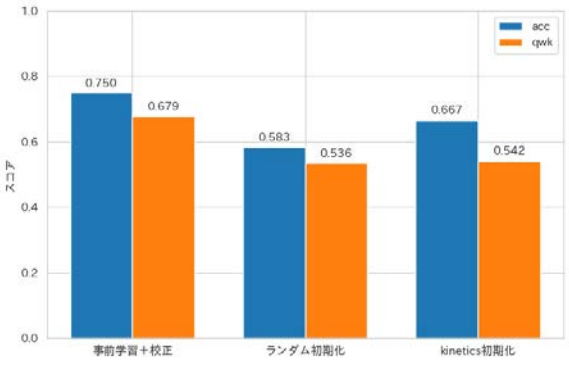



図 3: Accuracy と QWK (左)JIGSAWS 事前学習＋校正, (中央) ランダム初期化, (右)kinetics-400 で初期化

期化し，顕微鏡縫合動画のみで学習

### 4.4 時系列スコア表示

また，本研究では動画スコアの分類に加え，スライディングウィンドウで取得されるクリップの推論を並べることで得られる熟練度の時系列スコア（3.5 節）をヒートマップとして可視化し，評価の妥当性と教育的有用性を検討するために熟練医師によるレビューを実施した．評価者は 1 名の熟練医師であり，初心者 2 本・中級者 4 本・上級者 2 本の計 8 本の動画について，自由記述によるレビューを得た．

## 5. 結果

### 5.1 総合結果（Accuracy / QWK）

3 条件での Accuracy・QWK を図 3 にまとめた．提案法（JIGSAWS 事前学習＋校正）は Accuracy=0.750, QWK=0.679 を達成し，Kinetics-400 初期化（Accuracy=0.667, QWK=0.542）および ランダム初期化（Accuracy=0.583, QWK=0.536）を上回った．提案法は ランダム初期化 に対して Accuracy で+0.167, QWK で+0.143 の改善，Kinetics-400 初期化に対して Accuracy で+0.083, QWK で+0.137 の改善を示した．

### 5.2 混同行列とクラス別の傾向

図 4 に各条件の混同行列を示す．提案法は対角成分が全体に均整で，とくに上級（2），中級（1）での正解率が高い（約 0.78）．一方で Kinetics-400 初期化では初級→中級への取り違え（約 0.50）と上級→中級への取り違え（約 0.33）が目立ち，ランダム初期化では上級→中級が顕著（約 0.67）．すなわち，中級への取り違えが生じやすいバイアスが生じ，上下クラスの識別境界が曖昧になりやすいことが確認できる．

### 5.3 時系列ヒートマップ

代表例を図 5 に示す．分布の色は，赤が初心者，緑が中級者，青が上級者に対応している．動画全体を通じた色分布には，ラベルとおおよそ整合する傾向が観察された（初心者例では赤が多く，中級例では緑が多い）．一方で，上級者の動画でも青は限定的で，緑が優勢となる場面が多かった．青になる場面は一つの縫合が終わり場所を移すためガーゼを動かす場面，針を繊維に通し引っ張る場面に多かったが，上級判定の持続性が十分に反映されていない可能性が示唆された．

#### 5.3.1 医師レビュー

時系列可視化に対する医師レビューでは，一部の動画で可視化と臨床的判断が一致する例が確認されたものの，大半の動画でズレが報告された．具体的には，(i) 器具がフレームアウトし画面に有意な動作が映っていない時間帯でも緑（中級）として扱われる，(ii) 動作と関係の薄いクリップが評価対象になっている，(iii) 医師が重視する針先の角度・把持の安定性・糸のテンションといった微細運動に対するモデルの追従が不十分，の 3 点が見受けられた．これらは，現状の動画全体からの均等サンプリングと平均集約が，非動作区間や背景の影響を十分に排除できていないことに起因する可能性が高い．

## 6. 考察

### 6.1 本研究で得られた要点

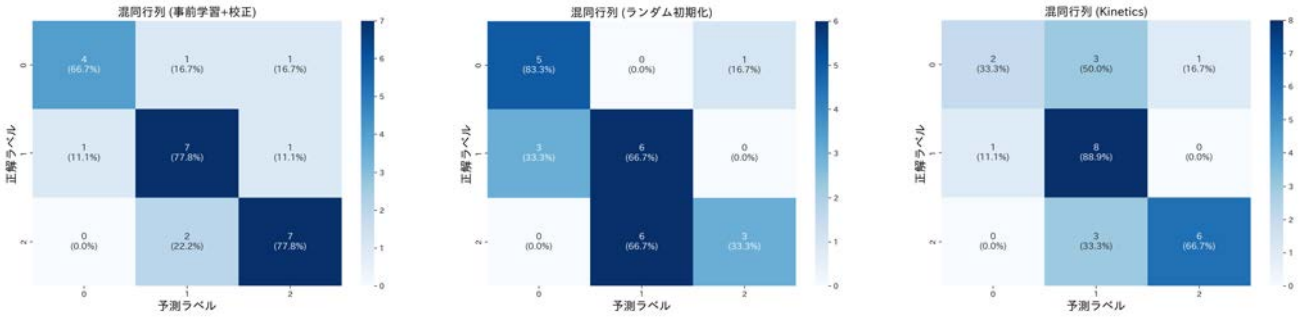
提案法（JIGSAWS 事前学習＋校正）は, 比較条件（Kinetics 初期化，ランダム初期化）に対して Accuracy と QWK の両方で優位だった．被験者独立（LOUO）でも一貫して差が維持された点は，単なる過学習ではなく決定境界の適正化が効いていることを示唆する．

### 6.2 校正（アイソトニック＋閾値最適化）が効いた理由

本タスクはドメインギャップ（内視鏡ロボット動画での事前学習 → 顕微鏡縫合）と熟練度クラスの順序性を併せ持つ．アイソトニック回帰は単調性の維持とスケール整合を同時に満たすため，外部事前学習モデルの出力を対象ドメインのラベル分布に寄せることができる．さらに，QWK 最大化で 2 閾値 $t_0, t_1$ を学習データから求めることで境界の位置が分布実態に適応し，中級へ寄りがちな誤り（5.2 節）を抑制できたと解釈できる．

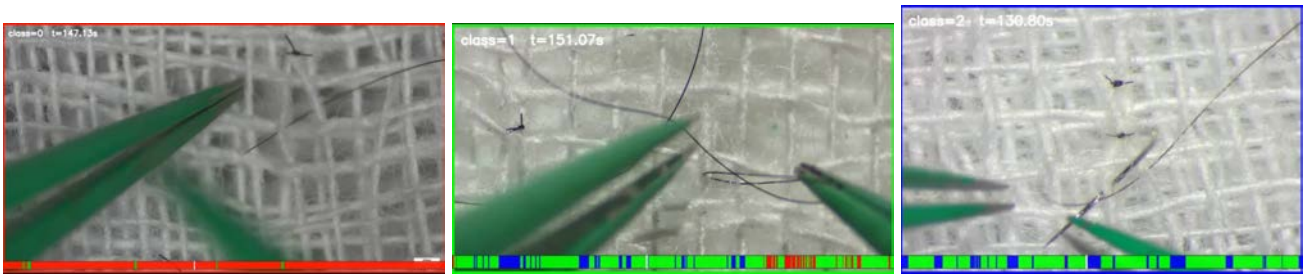
### 6.3 初期化の違い：Kinetics vs ランダム

本研究の比較では，Kinetics-400 で事前学習された重みで初期化したモデルが，ランダム初期化に比べて動画レベルの Accuracy / QWK が一貫して高い傾向を示した．この結果は，事前学習により汎用的な時空間特徴抽出能力が獲得され，限られた顕微鏡縫合データでも効率的に最適化が進むことを示唆する．具体的には，エッジ・運動・器具形状などの低〜中レベル表現がすでに整っているため，学習がラベル境界の調整やタスク固有の差異に早期に集中できると考えられる．




(a) JIGSAWS 事前学習＋校正 (提案)　(b) ランダム初期化　(c) Kinetics-400 で初期化

図 4: 3 つの混同行列（ラベルは 0: 初心者，1: 中級者，2: 上級者）



(a) 初心者　(b) 中級者　(c) 上級者

図 5: レベル別時系列ヒートマップ例（赤: 初心者，緑: 中級者，青: 上級者）

### 6.4 時系列評価

ヒートマップはおおまかな傾向を可視化するには有用だが，上級者でも青が乏しく緑が目立ち，医師評価との乖離が残った．これは，(1) 視覚的に目立つが技能とは無関係な変化（位置直し）を拾いやすい，(2) 技能判断に本質的な微細運動（針先の回転・刃先の入り方・テンション制御）がフレーム全体の表現に埋もれる，(3) 非動作区間の扱いが未分離，という三層の問題に整理できる．現行の時系列可視化は，学習支援の手掛かりとして有望とは言いにくい．中級者へ向けた導入的手掛かりとしては有用な可能性もあるが，微細運動と非動作区間の扱いに課題が残る．また，今回のレビューは評価者 1 名・自由記述のみの評価であり，評価者間一致や一般化可能性を検証するには不十分である．今後は複数評価者による共通プロトコル（選択式質問＋自由記述，あるいは区間単位の妥当性判定）を整備し，可視化が示す要改善区間の妥当性を定量化する必要がある．

### 6.5 現状の評価

本研究の目的は，学習者へのフィードバック提示（時系列可視化）を通じた，訓練支援である．この用途では，動画全体の自動採点に加えて，(i) 同一動画内での上手さの相対的な山/谷が再現されること，(ii) 少数データ下でも被験者独立で一定の一貫性を維持できること，が重要となる．一方で，本研究で得られた Accuracy=0.750 および時系列可視化の結果は，現時点で訓練支援へ直ちに適用できる水準を主張するものではない．しかし，提案法はランダム初期化および Kinetics 初期化よりも性能が高く，少数データ環境でも転移学習の有効性が示された．また医師レビューにより，推論時の非動作区間の混入や把持安定性・糸テンションといった微細運動の追従不足など，可視化が破綻する要因が具体的に指摘された．したがって今後は，フェーズ分類による非動作区間の除外と，器具や針先に注目した関心領域（ROI）の導入を組み合わせることで，推定精度と可視化の妥当性を実用に近い水準へ改善できる可能性がある．

## 7. まとめ

本研究では，脳神経外科の顕微鏡縫合訓練における訓練支援を目的として，ロボット支援手術動画からの転移学習を用いた顕微鏡縫合熟練度推定と時系列可視化の枠組みを提案した．JIGSAWS データセットで事前学習したモデルを顕微鏡縫合タスクへ転移し，アイソトニック回帰と閾値最適化によりスコアを校正した結果，Accuracy=0.750，QWK=0.679 を達成し，Kinetics 初期化およびランダム初期化を上回った．さらに，動画中の時系列スコアをヒートマップとして提示し，熟練医師（1 名）による自由記述レビューから，非動作区間の混入や微細運動の追従不足といった課題を含みつつも，可視化の有用性と改善方向（非動作区間の除外，ROI 導入等）を具体化した．以上より，

外部データによる時空間表現の転移と単調校正の併用は，限られた医療訓練データにおいて技能推定性能を改善し，学習者へのフィードバック提示へ接続するための有望な基盤となり得ることを示した．

**謝辞** 本研究は，JST さきがけ JPMJPR23I9 の支援を受けたものです．

## 参考文献

[1] 井上智弘，國井尚人，　熊切敦，大谷亮平，　田村晃，　齋藤勇，　堤一生：脳卒中外科手術技量の継承における卓上型マイクロによる縫合練習の役割―8 万針の効果―，脳卒中の外科，Vol. 37, No. 4, pp. 247–252 (2009).

[2] Tashiro, Y., Miyafuji, S., Kojima, Y., Kiyofuji, S., Kin, T., Igarashi, T. and Koike, H.: MR Microsurgical Suture Training System with Level-Appropriate Support, *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, Association for Computing Machinery, (online), DOI: 10.1145/3613904.3642324 (2024).

[3] Kojima, Y., Miyafuji, S., Tashiro, Y., Kiyofuji, S. and Koike, H.: MR MANE: MR Microsurgical Suturing Skill Acquisition for Novice Using Imitation of Example, *Proceedings of the 2024 International Conference on Advanced Visual Interfaces*, AVI '24, New York, NY, USA, Association for Computing Machinery, (online), DOI: 10.1145/3656650.3656660 (2024).

[4] De Mauro, A., Raczkowsky, J., Halatsch, M. E., HeinzW ¥"orn: Mixed Reality Neurosurgical Microscope for Training and Intra-operative Purposes, *Virtual and Mixed Reality* (Shumaker, R., ed.), Berlin, Heidelberg, Springer Berlin Heidelberg, pp. 542–549 (2009).

[5] Tashiro, Y., Miyafuji, S., Hwang, D.-H., Kiyofuji, S., Kin, T., Igarashi, T. and Koike, H.: GAuze-MIcrosuture-FICATION: Gamification in Microsuture training with real-time feedback, *Proceedings of the Augmented Humans International Conference 2023*, AHs '23, New York, NY, USA, Association for Computing Machinery, p. 15–26 (online), DOI: 10.1145/3582700.3582704 (2023).

[6] Eom, S., Kim, S., Rahimpour, S. and Gorlatova, M.: AR-Assisted Surgical Guidance System for Ventriculostomy, (online), DOI: 10.1109/VRW55335.2022.00087 (2022).

[7] Sharma, R. and Suri, A.: Microsurgical suturing assessment scores: a systematic review, *Neurosurgical Review*, Vol. 45, No. 1, pp. 119–124 (2022).

[8] Martin, J., Regehr, G., Reznick, R., Macrae, H., Murnaghan, J., Hutchison, C. and Brown, M.: Objective structured assessment of technical skill (OSATS) for surgical residents, *British journal of surgery*, Vol. 84, No. 2, pp. 273–278 (1997).

[9] Reznick, R., Regehr, G., MacRae, H., Martin, J. and McCulloch, W.: Testing technical skill via an innovative "bench station" examination, *The American journal of surgery*, Vol. 173, No. 3, pp. 226–230 (1997).

[10] Wan, B., Peven, M., Hager, G., Sikder, S. and Vedula, S. S.: Spatial-temporal attention for video-based assessment of intraoperative surgical skill, *Scientific reports*, Vol. 14, No. 1, p. 26912 (2024).

[11] Gao, Y., Vedula, S. S., Reiley, C. E., Ahmidi, N., Varadarajan, B., Lin, H. C., Tao, L., Zappella, L., Béjar, B., Yuh, D. D. et al.: Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling, *MICCAI workshop: M2cai*, Vol. 3, No. 2014, p. 3 (2014).

[12] Wang, Z. and Majewicz Fey, A.: Deep learning with convolutional neural network for objective skill evaluation in robot-assisted surgery, *International journal of computer assisted radiology and surgery*, Vol. 13, No. 12, pp. 1959–1970 (2018).

[13] Lajkó, G., Nagyne Elek, R. and Haidegger, T.: Endoscopic image-based skill assessment in robot-assisted minimally invasive surgery, *Sensors*, Vol. 21, No. 16, p. 5412 (2021).

[14] Raghu, M., Zhang, C., Kleinberg, J. and Bengio, S.: Transfusion: Understanding transfer learning for medical imaging, *Advances in neural information processing systems*, Vol. 32 (2019).

[15] Cheplygina, V., De Bruijne, M. and Pluim, J. P.: Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis, *Medical image analysis*, Vol. 54, pp. 280–296 (2019).

[16] Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y. and Paluri, M.: A closer look at spatiotemporal convolutions for action recognition, *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 6450–6459 (2018).

[17] Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P. et al.: The kinetics human action video dataset, *arXiv preprint arXiv:1705.06950* (2017).

[18] Tong, Z., Song, Y., Wang, J. and Wang, L.: Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training, *Advances in neural information processing systems*, Vol. 35, pp. 10078–10093 (2022).

[19] Zhang, D., Wu, Z., Chen, J., Gao, A., Chen, X., Li, P., Wang, Z., Yang, G., Lo, B. and Yang, G.-Z.: Automatic microsurgical skill assessment based on cross-domain transfer learning, *IEEE Robotics and Automation Letters*, Vol. 5, No. 3, pp. 4148–4155 (2020).

[20] Tang, Y., Ni, Z., Zhou, J., Zhang, D., Lu, J., Wu, Y. and Zhou, J.: Uncertainty-aware score distribution learning for action quality assessment, *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9839–9848 (2020).

[21] Dong, X., Liu, X., Li, W., Adeyemi-Ejeye, A. and Gilbert, A.: Interpretable long-term action quality assessment, *arXiv preprint arXiv:2408.11687* (2024).

[22] Wang, T., Wang, Y. and Li, M.: Towards accurate and interpretable surgical skill assessment: A video-based method incorporating recognized surgical gestures and skill levels, *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, pp. 668–678 (2020).

[23] Chang, S., Liu, D., Tian, R., Swartzell, K. L., Klingler, S. L., Nagle, A. M. and Kong, N.: Automated Procedural Analysis via Video-Language Models for AI-assisted Nursing Skills Assessment, *arXiv preprint arXiv:2509.16810* (2025).

[24] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L. et al.: Pytorch: An imperative style, high-performance deep learning library, *Advances in neural information processing systems*, Vol. 32 (2019).

[25] Cao, W., Mirjalili, V. and Raschka, S.: Rank consistent ordinal regression for neural networks with application to age estimation, *Pattern Recognition Letters*, Vol. 140,

pp. 325–331 (2020).
[26] Carreira, J. and Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset, *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308 (2017).