Vol.2023-CVIM-232 No.10
2023/1/26

撮影順序情報を活用した潰瘍性大腸炎分類モデルの提案

原田　翔太[1,a)]　備瀬　竜馬[1]　田中　聖人[2]　内田　誠一[1]

概要：潰瘍性大腸炎 (UC) の分類は内視鏡診断の重要な課題の 1 つであるが，主に 2 つの困難性がある．第一の困難性は，UC ラベル (正常または炎症) 付き内視鏡画像の枚数が制限される点である．第二の困難性は，内視鏡画像が捉えている臓器の場所が変化に従い，その画像の概観も大きく変化する点である．特に，第二の困難は，既存の半教師付き学習手法の適用を妨げている．本研究では，内視鏡画像から容易に取得可能な 2 種類の情報，臓器内の場所情報 (例：左結腸) と内視鏡画像の画像の撮影順序を利用する，UC 分類のための新たな半教師付き学習手法を提案する．さらに，内視鏡画像の系列内のサンプル間の関係性をより柔軟に捉えるために，Transformer を導入した画像系列分類モデルも提案する．実験結果から，提案手法は既存の半教師付き学習法よりも分類精度の改善に有効であることが確認された．

1. はじめに

深層学習を用いた潰瘍性大腸炎 (UC) 分類において，内視鏡画像に対するアノテーションを実施するには専門医の協力が必要なため，ラベル付きサンプルを大量に収集することは困難である．UC は大腸に炎症と潰瘍を引き起こす炎症性腸疾患である．UC の診断に重要となる所見は血管透見像の有無や潰瘍の度合いなどの微細な特徴であり，画像からそれらの所見の有無を判断し，UC ラベルをアノテーションするためには医学的な専門知識が必要である．

半教師付き学習はラベル付きサンプルが限られた状況で，高精度な分類モデルを獲得するための学習手法であり，多くの研究成果が報告されている [1], [2], [3]．少数のラベル付きサンプルによって学習された分類器の推定結果がある程度信頼できる場合，既存の半教師付き学習手法による分類性能の向上は期待できる．しかし，内視鏡画像を正常と炎症の 2 クラスに分類する問題 (UC 分類) の場合，既存手法を単に適用しただけでは良い効果を得ることは難しい．なぜならば，既存手法の多くは画像の主要な概観は分類対象のクラスに依存していることを暗黙的に仮定しているのに対して，内視鏡画像の場合，UC に依存した特徴は非常に微細であり，内視鏡画像の主要な概観は臓器の場所に依存するためである．

ラベル付きサンプル不足を補う別の方策として，対象タスクのドメイン知識を学習に導入する方法がある．この方策は，分類モデルの推定結果が信頼できない場合でも，有効に動作する可能性がある．例えば，内視鏡画像においては，臓器の場所情報と内視鏡画像の撮影順序情報は容易に取得できる．臓器の場所情報（臓器の場所ラベル）は検査中の内視鏡の動きを追跡することで，比較的容易に取得可能である [4], [5]．また，内視鏡画像は内視鏡画像系列として撮影されるため，撮影順序情報も容易に取得できる．そして撮影する際には，内視鏡を臓器の中を移動させながら撮影していく．そのため，時間隣接したサンプルは同一の UC クラスに属している可能性が高い．本研究ではこれらの情報を活用した新たな学習方法と分類モデルを提案する．

本研究では，UC 分類のための半教師付き学習手法として内視鏡画像の場所ラベルと撮影順序情報を活用する Order-guided disentangled representation learning を提案する．本研究では，内視鏡画像系列内の一部サンプルに UC ラベルのアノテーションがされている状況を想定しており，場所ラベルと順序情報に関しては全サンプルで利用可能な設定となっている．

提案手法では場所ラベルを用いた Disentangled representation learning[6], [7] を導入することで，臓器の場所に依存しない特徴 (UC に依存した特徴) を抽出することを目指す．Disentangled representation learning は複数の要因に依存した特徴量を各要因にのみ依存した特徴へと分離することを目的とした学習手法である．本手法では，UC ラベルと場所ラベルを利用した Disentangled representation learning によって場所情報に依存しない UC 特徴を獲得することを目指す．

内視鏡画像系列の撮影順序情報を用いる Order-guided learning も提案する．この学習方法は，時間隣接したサン

[1] 九州大学
[2] 京都第二赤十字病院
[a)] shota.hrada@human.ait.kyushu-u.ac.jp

プルのペア (同一の UC クラスに属する可能性の高いサンプルのペア) を特徴空間上で近づけるようにモデルを最適化する．この学習手法は UC ラベル無しサンプルに対しても適用可能であるため，UC ラベル付きサンプル不足を解決できる可能性がある．

さらに本研究では，内視鏡画像系列内の関係性をより柔軟に取り扱うための分類モデル (画像系列分類モデル) を提案する．前述の手法では，時間隣接サンプルの関係性を画像分類モデルの学習に利用していたが，内視鏡画像系列内には時間隣接サンプル間の関係性以外にも UC 分類に重要な情報が含まれると考えられる．本研究では，系列内の関係性を柔軟に取り扱うことを可能にするために，Transformer を導入した画像系列分類モデルを提案する．

本研究の主な貢献は以下の通りである．

- 内視鏡画像の場所情報と撮影順序情報を活用した UC 分類のための半教師付き学習手法の提案
- 内視鏡画像の撮影順序情報を学習に導入するための学習方法 (Order-guided learning) の提案．
- 内視鏡画像系列内のサンプル間の関係性をより柔軟に捉えることを目的とした，Transformer ベースの画像系列分類モデルの提案

2. 先行研究

ラベル付きサンプルが制限されている状況における分類器の学習において，ラベル無しサンプルを精度改善のために有効活用する半教師付き学習手法が数多く提案されている [1], [2], [3], [8]．Lee らは，少量のラベル付きサンプルで学習された分類器の予測結果をラベル無しサンプルに対して疑似ラベルとして与えて学習する Pseudo-Label を提案した [2]．本手法は単純な方法であるのにも関わらず，いくつかのデータセットに対して，精度が改善することが報告されている．Song らは，弱いデータ拡張を施したラベル無しサンプルと，強いデータ拡張を施した同一のラベル無しサンプルに対する予測結果を一致させるように学習する，FixMatch を提案している [3]．

これらの手法は，少量のラベル付きサンプルで学習された分類器の性能がある程度高い場合，良い効果が期待できる．しかし，血管透見像の有無や潰瘍の度合いといった微細な特徴が重要となる UC 分類においては，少量のラベル付きサンプルのみから性能の高い分類器を獲得することは困難である．そのため，これらの既存手法をそのまま UC 分類に適用することは困難である．

動画像などの画像系列の時間順序情報を活用した学習手法はいくつか提案されている [9], [10]．例えば，Cao らは系列間の対応付けを活用した自己教師あり学習，Temporal-Cycle Consistency (TCC) を提案している．TCC は系列間の一貫性のある対応付けが容易になるような特徴空間を獲得するように学習する．Dwibedi らは，ラベル付き動画像とラベル無し動画像の時間方向での対応付けを利用する学習方法を提案した [10]．この方法は，時間方向で対応付けられたフレーム間の距離を最小化するように学習することで，動画像分類の精度を向上している．これらの手法は，系列内のクラス遷移が同じであることを仮定している．本研究で対象としている UC 分類は系列内の UC クラスの遷移が不定であるため，これらの手法を適用することはできない．

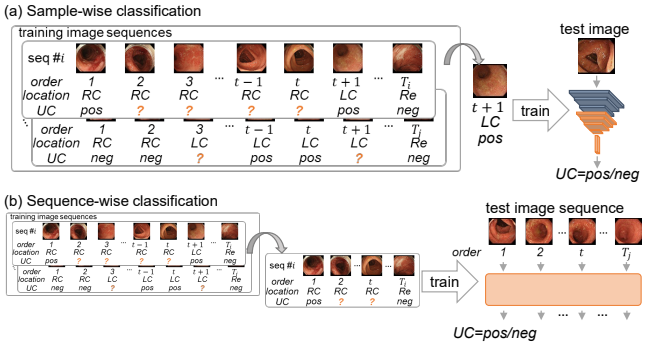

図 1 (a) 画像分類モデルと (b) 画像系列分類モデルの概要

3. 内視鏡画像の事前知識を活用した画像分類モデル

図 1(a) に画像分類モデルの概要を示す．画像分類モデルは画像単位で入力を受け取り，画像毎に UC クラスを推定する．本研究では画像分類モデルを学習するために，臓器の場所ラベルと撮影順序情報を活用した学習方法を提案する．この学習方法により，UC ラベル付きサンプル数が制限されている状況でも高精度な分類モデルが獲得することが期待される．

画像分類モデルの学習方法として，内視鏡画像の場所情報と撮影順序情報を活用した Order-guided disentangled representation learning を提案する．提案手法は，Disentangled representation learning と Order-guided learning の 2 つから構成されている．臓器の場所情報 (場所ラベル) は UC ラベルに比べて比較的取得しやすい．そこで，臓器の場所ラベルを利用して Disentangled representation learning することで，画像特徴を場所特徴および UC 特徴に分離する．Order-guided learning は内視鏡画像の撮影順序を活用した学習方法であり，内視鏡画像系列中の時間的に隣接するサンプルは同一の UC クラスに属する傾向があるという特性を利用する．この特性を活用した目的関数を定式化し，UC 分類モデルの学習に用いることで，UC ラベル付き画像の制限を解決する．

3.1 場所ラベルを活用した Disentangled representation learning

提案手法のネットワーク構造を図 2(a) に示す．このネットワークは，特徴抽出器 E_{enc} からの出力が 2 つのタスク

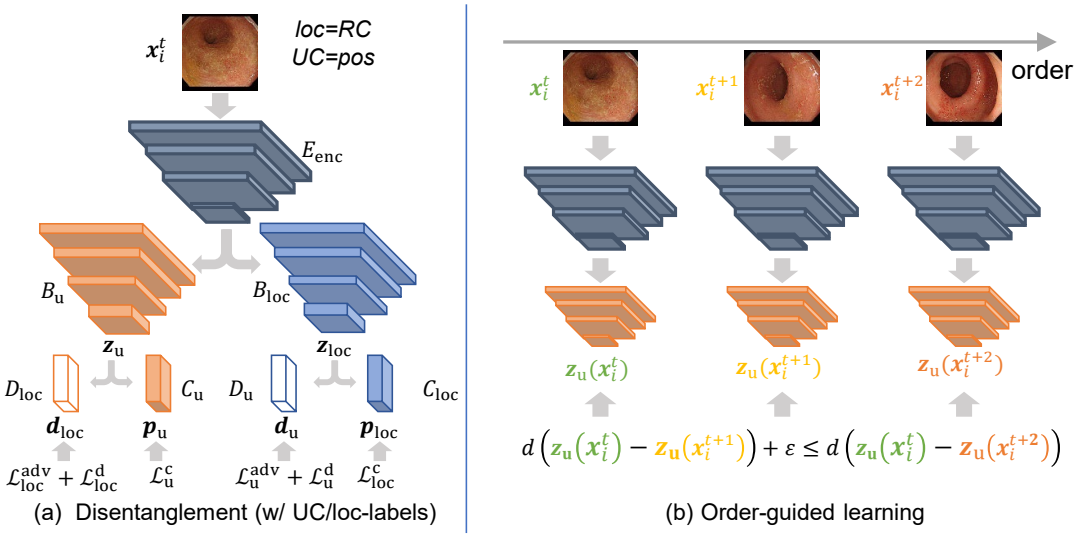

図 2 画像分類モデルの詳細. (a)UC 特徴 $\boldsymbol{z}_{\mathrm{u}}$ と場所特徴 $\boldsymbol{z}_{\mathrm{loc}}$ を分離するための Disentanglement representation learning. (b) 撮影順序情報を活用する Order-guided learning.

固有の特徴抽出器 $B_{\mathrm{u}}, B_{\mathrm{loc}}$ に分岐し，各モジュールがさらに 2 つの分類層に分岐する階層構造である．特徴抽出器 E_{enc} は，入力画像から UC クラスと場所クラスを分類するための共通特徴ベクトルを抽出する．タスク別の特徴抽出器 B_{u} と B_{loc} は，UC と場所の分類のために分離された特徴である UC 特徴 $\boldsymbol{z}_{\mathrm{u}}$ と場所特徴 $\boldsymbol{z}_{\mathrm{loc}}$ を抽出する．

4 つの分類層の内，C_{u} と C_{loc} は，それぞれ UC 分類と場所分類に用いられ，これらの出力は一般的な分類損失に基づいた学習に利用される．一方，D_{u} と D_{loc} は Disentangled representation learning のために利用される．Disentangled representation learning のための目的関数はいくつか提案されているが，本手法では Liu らが提案しているクラス予測確率のエントロピーの最大化を活用した敵対的学習に従ってモデルを学習する [7].

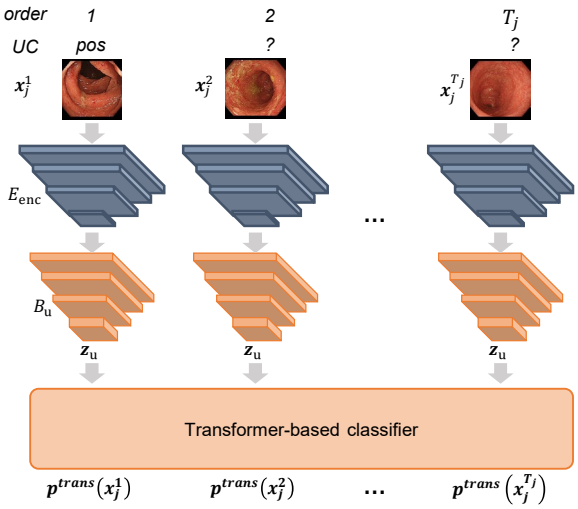

図 3 画像系列分類モデルの構造

3.2 Order-guided learning

Order-guided learing の概要を図 2(b) に示す．図 2(b) に示すように，Order-guided learning は時間隣接した画像を入力しそれらの特徴量を近づける学習方法である．Order-guided learning は下式で定義する．

$$\mathcal{L}_{\mathrm{seq}}(\boldsymbol{x}_i^t, \boldsymbol{x}_i^{t+1}, \boldsymbol{x}_i^{t+2}) = \left[||\boldsymbol{z}_{\mathrm{u}}(\boldsymbol{x}_i^t) - \boldsymbol{z}_{\mathrm{u}}(\boldsymbol{x}_i^{t+1})||_2^2 - ||\boldsymbol{z}_{\mathrm{u}}(\boldsymbol{x}_i^t) - \boldsymbol{z}_{\mathrm{u}}(\boldsymbol{x}_i^{t+2})||_2^2 + \varepsilon \right]_+ . \quad (1)$$

ここで，$\boldsymbol{z}_{\mathrm{u}}(\boldsymbol{x}_i^t)$ はサンプル $\boldsymbol{x}_i^t$ の UC 特徴ベクトルで，E_{enc} と B_{u} を通じて抽出される．また，$[\cdot]_+$ は負の入力に対して 0 を返し，それ以外は入力を直接出力する関数，ε は時間的に離れた 2 つのサンプル間の不一致の度合いを制御するマージンである．

Order-guided learning によってモデルを最適化することで，時間的に隣接するサンプル間の UC 特徴が時間的に離れたサンプルの UC 特徴よりも近づく．つまり，同一の UC クラスに属している可能性が高いサンプルが特徴空間上で近づくことになる．その結果，限られた UC ラベル付きサンプルのみで学習するよりも，高精度な分類モデルを獲得できることが期待される．

4. Transformer を活用した画像系列分類モデル

図 1(b) に画像系列分類モデルの概要を示す．提案モデルは入力として画像系列を受け取り，系列内の各サンプルの UC クラスを一括で予測する．画像系列分類モデルは，3 章で紹介した画像分類モデルを利用してサンプル毎に特徴を抽出する．その後，抽出された特徴の系列を Transoformer に入力し，系列内の各サンプルの UC ラベルを一括して推定する構造となっている．学習を通して自動的に UC 分類に有用な系列内の関係性を捉えることを期待して，Transformer は導入されている．

4.1 内視鏡画像系列内の関係性

内視鏡画像は医師が内視鏡を臓器内で動かしながら撮影されるため，実際には画像系列として取得される．この画像系列内には，3.2 章の Order-guided learning の基になっている UC ラベルの時間連続性をはじめとした，個別のサンプルからは獲得できない UC 分類に有用な情報が含まれていると考えられる．具体的には，内視鏡画像系列からは，1) 系列内のサンプル間の依存関係，2) 画像系列内の共通特徴，の 2 種類の情報が獲得できると考えられる．

時間隣接サンプル間の依存関係：内視鏡画像の UC ラベルは時間連続性を持つと考えられる．前述の通り，内視鏡画像は内視鏡を臓器内で移動させながら撮影される．そのため，撮影順序が隣接したサンプルは臓器内の空間的に近い箇所を捉えている可能性が高い．つまり，撮影順序が近いサンプルは同一の UC クラスに属している可能性が高いということである．

時間隣接したサンプル間の関係性はは Order-guided learning でも利用しているが，この学習方法では 1 フレーム隣のサンプル間の関係性のみの利用に限られる．しかし，実際には 1 フレーム以上隣のサンプルであっても同一の UC クラスに属する可能性はある．画像系列分類モデルでは，画像系列を入力することで，学習を通じて撮影順序が近いサンプル間の関係性を柔軟に獲得することを目指す．

また，撮影順序が遠いサンプルに関しても臓器内の空間的に近い箇所を捉えている可能性がある．内視鏡検査では，内視鏡を観察対象の臓器の最奥まで挿入し，抜去しながら観察される．一般的には詳細な観察は抜去時に実施されるが，挿入時にも観察される．そのため，撮影順序が隣接していなくても，同一の臓器や疾病を捉えている可能性がある．

遠方サンプルのペア候補は多くのパターンが考えられるため，遠方サンプル間の関係性を利用した学習を Order-guided learning のようなサンプルのペアを入力とした表現学習で定義することは難しい．一方，画像系列分類モデルでは，画像系列を一括して入力するため，仮に遠方サンプル間の関係性が UC 分類に有用である場合，それらのサンプルの関係性を積極的に活用することが有用であるということを，学習を通して自動的を獲得することが期待できる．

画像系列内の共通特徴：画像系列全体からは，患者個人の特徴の獲得が期待できる．臓器の概観は個人ごとに色や形状が異なる．さらに，撮影時の光源の強さや内視鏡の種類などの条件の変化に従い，画像の概観が変化することもある．このような特徴は実際の UC 分類において不要である．画像系列分類モデルでは画像系列全体を入力することで，系列内で共通しているような特徴には依存しない特徴を抽出可能になることが期待される．

4.2 画像系列分類モデルの詳細

提案モデルは Transformer を導入することで，系列内から UC 分類に重要なサンプル間の関係性を獲得することを目指す．Transformer の内部にある Self-Attention を活用することで，上記の 3 種類の関係の全てを表現できると考えた．図 3 にモデルの概要を示すように，図 3 の画像分類モデルの分類層 C, D を Transformer ベースの分類器に置き換えた構造となっている．

提案モデルでは，撮影順序情報を画像系列に埋め込むために，Positional Encoding を導入する．今回は，絶対位置表現に基づいて情報を埋め込む Absolute Position Encoding (APE) [11] を採用する．通常，APE は入力系列の長さで正規化しないが，本手法は内視鏡画像系列が対象のため，系列の長さで正規化した値を利用して位置情報を埋め込む．

画像系列分類モデルは 2 ステップで学習する．最初に Transformer 以外を 3 章で紹介した画像分類モデルの学習方法（Order-guided disentangled representation learning）に従って事前学習する，その後，UC 分類のためのクロスエントロピーに基づいた分類損失を目的関数として画像系列分類モデル全体を再学習する．事前に Disnentangled representation learning によって UC 依存の特徴のみを抽出することで，系列内のサンプル間の関係性の抽出が容易になると考えた．

5. 実験結果

5.1 データセット

本実験で利用したデータセットは 388 の内視鏡画像系列が含まれ，各系列の長さは異なる．また，データセット中のサンプル数は 10,262 である．各サンプルは専門医によって UC ラベルと臓器の場所ラベルがアノテーションされている．UC ラベルは正常と炎症の 2 クラス，場所ラベルは右結腸，直腸，そして左結腸の 3 クラスでアノテーションされている．10,262 サンプルの内，6,678 サンプルが正常，5,584 サンプルが炎症としてアノテーションされた．実験のために，データセットは画像系列基準で，232，77，79 に分割され，それぞれを学習セット，検証セット，テストセットとして利用した．また，半教師付き学習設定にするため，学習セットから 10%のサンプルをランダムに抽出し，それらを UC ラベル付きサンプルとして扱った．

5.2 比較手法

本実験では 2 種類の半教師あり学習方法と提案手法を比較した．第一の方法は，有名な半教師付き学習法の 1 つである，疑似ラベル法である [2]．第二の手法は，一般画像を対象とした半教師付き分類タスクで良く用いられる FixMatch[3] である．ベースラインとして学習セット中の全サンプルを UC ラベル付きサンプルとして扱い学習し

表 1 分類精度の比較．Sample-wise model は画像分類モデル，Sequence-wise model は画像系列分類モデルを表す．R は学習セット中の UC ラベルを与えたサンプルの割合を示す．最高精度は**太字**で示し，2 番目に優れた精度には下線を引いている．

Method	R	Precision	Recall	Specificity	Accuracy	F1
Fully-supervised learning	1.0	0.8052	0.8489	0.9023	0.8851	0.8264
Supervised learning	0.1	0.6916	0.6707	0.8578	0.7975	0.6810
Pseudo-Label [2]	0.1	0.7519	0.6133	0.9037	0.8110	0.6755
FixMatch [3]	0.1	**0.7524**	0.4682	**0.9267**	0.7790	0.5773
Sample-wise model	0.1	0.7335	0.8018	0.8270	0.8176	0.7661
Sequence-wise model	0.1	0.7331	0.8255	0.8215	0.8230	0.7765
Sequence-wise model w/o APE	0.1	0.7450	0.8281	0.8316	**0.8303**	**0.7843**
Sequence-wise model w/ single	0.1	0.7234	**0.8307**	0.8114	0.8186	0.7734

表 2 画像分類モデル (Sample.) と画像系列分類モデル (Sequence.) のサンプル毎の正答/誤答の変化

		Sequence.	
		Correct	Incorrect
Sample.	Correct	1,617	55
	Incorrect	81	292

た分類モデル (Fully-supervised learning) と 10%の UC ラベル付きサンプルのみを利用して学習された分類モデル (Supervised learning) と比較した．また，Ablation study として，画像系列分類モデルへ 1 サンプル毎に入力した画像系列分類モデル (Sequence-wise model w/ single) と Positional Encoding を導入していないモデル (Sequence-wise model w/o APE) と比較した．

5.3 定量評価

表 1 に各モデルの分類精度を示す．表 1 から，提案手法が既存手法よりも分類性能が高いことが確認できる．特に Recall と F1 は大幅に改善された．また，入力が画像単位のモデル (Sample-wise model と Sequence-wise model w/ single) と画像系列単位のモデル (Sequence-wise model と Sequence-wise model w/o APE) の比較から，画像系列単位での入力は UC 分類の精度改善に繋がることが分かった．一方で，Sequence-wise model と比較して，Sequence-wise model w/o APE の精度が優れていた．このことから，位置情報埋め込みは UC 分類の精度改善に寄与しない可能性があることが確認された．定量的な精度評価の結果，Sequence-wise model w/o APE が最高精度であったため，これ以降の解析では Sequence-wise model w/o APE を対象としている．

表 2 に画像分類モデルと画像系列分類モデルのサンプル毎の正答/誤答の変化を示す．この結果から，画像系列分類モデルを利用することで 88 サンプルが正答に，55 サンプルが誤答に転じたことが分かる．画像系列分類モデルは，学習に利用するラベル数やサンプル数に変化させずに，入力データの形式を変化させただけで精度を改善した．この知見は，今後，UC 分類モデルを構築する際に有用になると考える．

5.4 定性評価

図 4 に画像分類モデルと画像系列分類モデルによる推定結果例を示す．図 4(a) の左から 2，5，10 番目のサンプルは正答に転じていることが確認できる．これは，入力を画像系列として取り扱ったことで，系列内の推定が容易なサンプルの影響を受けたためだと考える．実際に，図 4(a) の左から 2 番目のサンプルは，画像分類モデルでも正答していた左から 1 番目のサンプルに比較的類似しており，このサンプルの影響を受けて，正答に転じた可能性が高い．

図 5(a) に画像系列分類モデルの Attention matrix の例を示す．i 番目の行の各数値が i 番目のサンプルをクエリとした時の Attention となっている．この結果から，どのサンプルに対しても類似した Attention が得られていることが確認される．この結果と Attention の強度を照らし合わせると，Transformer の Self-Attention は画像系列内の正常サンプルの平均に類似したものを出力していると考えることが出来る．つまり，“画像系列内の共通特徴” を抽出することが UC 分類の精度改善に有用である可能性が示唆された．正常サンプルにより注目している理由は，患者毎の臓器の見た目の違いや撮影条件などの系列固有の特徴を炎症サンプルからは抽出しづらかったためだと考える．

図 5(b) に画像系列分類モデルの Tranformer を通す前後の特徴量の分布を示す．これらの特徴量は元の特徴量から t-SNE を利用して次元圧縮された特徴量である．図 5(b) から，Transformer を適用することで，系列内のサンプルが相対的に近くなったことが分かる．この結果から，Transformer には系列内のサンプルからある程度一貫した特徴を抽出する働きを持つことが確認された．

6. まとめ

本研究では，内視鏡画像のドメイン知識を活用した UC 分類のための半教師付き学習手法を提案した．さらに，内視鏡画像系列内のサンプル間の関係性を柔軟に捉えるために，Transformer を導入した画像系列分類モデルも提案し

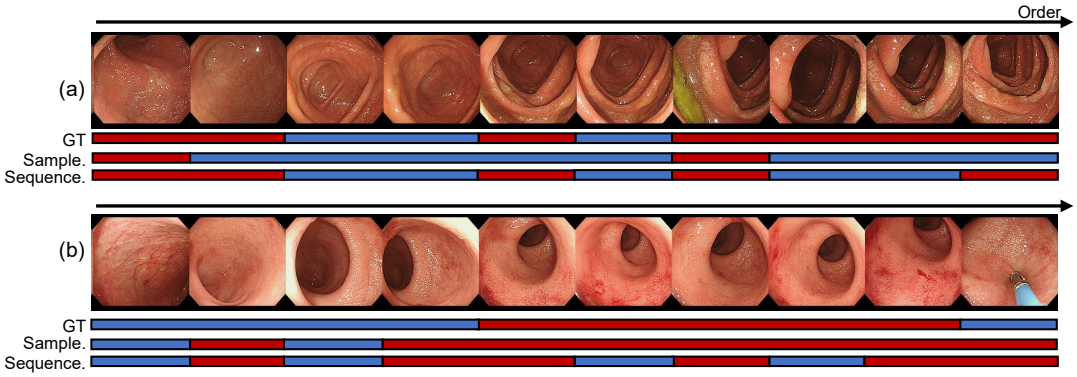

図 4 画像分類モデル (Sample.) と画像系列分類モデル (Sequence.) の推定結果例．(a) 画像系列分類モデルで精度が改善した例および (b) 改悪した例．(青：正常，赤：炎症)

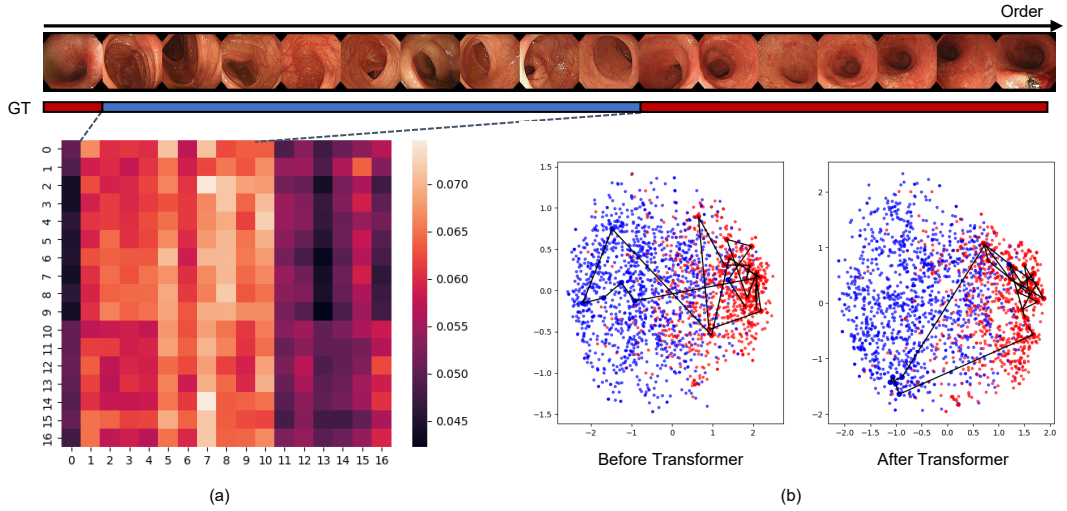

図 5 画像系列分類モデル (Sequence-wise model w/o APE) の (a)Attention matrix の例と (b)Transformer 適用前後の特徴量の分布 (青：正常，赤：炎症)．散布図中の黒線は時間隣接したサンプル間に与えられる．

た．実験結果から，提案手法は既存手法よりも高精度であることが確認された．また，画像分類モデルよりも画像系列分類モデルの精度が優れていたことから，UC 分類においては，画像系列毎に入力する方が適していること可能性が高いことも確認された．

謝辞：本研究は，JSPS 科研費 JP20H04211，JP21K18312，JP21J13083 の助成を受けたものである．

参考文献

[1] Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A. and Raffel, C. A.: MixMatch: A Holistic Approach to Semi-Supervised Learning, *NeurIPS*, Vol. 32 (2019).

[2] Lee, D.-H.: Pseudo-Label : The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks, *ICML Workshop* (2013).

[3] Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C. A., Cubuk, E. D., Kurakin, A. and Li, C.-L.: FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence, *NeurIPS*, Vol. 33, pp. 596–608 (2020).

[4] Herp, J., Deding, U., Buijs, M. M., Kroijer, R., Baatrup, G. and Nadimi, E. S.: Feature Point Tracking-Based Localization of Colon Capsule Endoscope, *Diagnostics*, Vol. 11, No. 2 (2021).

[5] Mori, K., Deguchi, D., Hasegawa, J.-i., Suenaga, Y., Toriwaki, J.-i., Takabatake, H. and Natori, H.: A Method for Tracking the Camera Motion of Real Endoscope by Epipolar Geometry Analysis and Virtual Endoscopy System, *MICCAI*, pp. 1–8 (2001).

[6] Liu, A. H., Liu, Y.-C., Yeh, Y.-Y. and Wang, Y.-C. F.: A Unified Feature Disentangler for Multi-Domain Image Translation and Manipulation, *NeurIPS*, Vol. 31 (2018).

[7] Liu, Y., Wei, F., Shao, J., Sheng, L., Yan, J. and Wang, X.: Exploring Disentangled Feature Representation Beyond Face Identification, *CVPR* (2018).

[8] Arazo, E., Ortego, D., Albert, P., O'Connor, N. E. and McGuinness, K.: Pseudo-Labeling and Confirmation Bias in Deep Semi-Supervised Learning, *IJCNN* (2020).

[9] Cao, K., Ji, J., Cao, Z., Chang, C.-Y. and Niebles, J. C.: Few-Shot Video Classification via Temporal Alignment, *CVPR* (2020).

[10] Dwibedi, D., Aytar, Y., Tompson, J., Sermanet, P. and Zisserman, A.: Temporal Cycle-Consistency Learning, *CVPR* (2019).

[11] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u. and Polosukhin, I.: Attention is All you Need, *NeurIPS*, Vol. 30 (2017).