

# 空間的特徴に着目した知識逆蒸留による 嚥下機能評価の検討

岩手大学  
理工学部 システム創成工学科  
知能・メディア情報コース  
鈴木 晴仁

令和8年2月27日

# 目 次

<b>第 1 章 緒論</b>	<b>1</b>
1.1 本研究の背景・目的	1
1.2 本論文の構成	3
<b>第 2 章 深層学習に基づく食塊検知法</b>	<b>4</b>
2.1 教師あり食塊検知法	4
2.2 教師なし食塊検知法	5
2.2.1 知識蒸留 (Knowledge Distillation)	6
2.2.2 知識逆蒸留 (Knowledge Reverse Distillation)	6
2.2.3 One-Class Bottleneck Embedding (OCBE)	8
2.2.4 知識蒸留損失 (KD Loss)	10
<b>第 3 章 注意機構を導入した知識逆蒸留に基づく食塊検知法の提案</b>	<b>11</b>
3.1 注意機構 (Attention Mechanism)	11
3.2 Convolutional Block Attention Module (CBAM)	12
3.2.1 Channel Attention Module	13
3.2.2 Spatial Attention Module	14
3.3 OCAE ブロックを用いた知識逆蒸留に基づく食塊検知法	15
3.3.1 Residual Network (ResNet)	16
3.3.2 Wide Residual Network (WRN)	17
3.3.3 残差ブロックに対する CBAM の導入	18
<b>第 4 章 評価実験</b>	<b>19</b>
4.1 実験設定	19
4.1.1 嘔下超音波画像データセット	19
4.1.2 評価指標	20
4.2 実験結果	21
4.3 分析	23
4.3.1 アブレーション研究	23
4.3.2 CBAM の導入位置に関する検討	24
4.3.3 CBAM の構成に関する検討	25
4.3.4 ハイパーパラメータおよびバックボーンネットワークに関する検討	26
<b>第 5 章 結論</b>	<b>28</b>
<b>謝辞</b>	<b>30</b>
<b>参考文献</b>	<b>33</b>

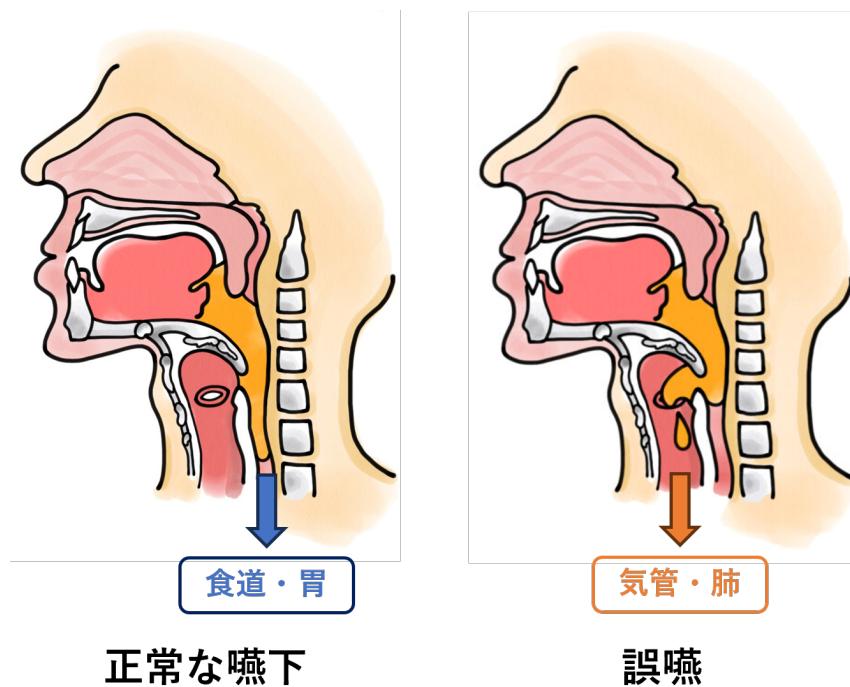
# 第1章 緒論

## 1.1 本研究の背景・目的

近年、高齢化による患者の増加や医師不足に伴い、AIを用いた医用画像処理が注目されている。特に、耳鼻咽喉科では嚥下障害患者の増加が著しく、臨床リハビリテーションにおける食塊検知モデルの需要が高まっている。

嚥下は飲食物を口腔から胃まで運搬する一連の運動であり、食塊が通過する器官とメカニズムにより以下3つの時期に分類される [1].

- i) 嚥下第1期（口腔期、随意期）  
食塊が舌運動により口腔から咽頭へ移動するまでの時期.
- ii) 嚥下第2期（咽頭期、反射期、不随意期）  
食塊が咽頭から食道入口部を通過するまでの時期.
- iii) 嚥下第3期（食道期、不随意期）  
食塊が食道入口部を通過した後に食道から胃へと送り込まれる時期.



estychan.comより改編

図 1.1: 嚥下障害の概要. 左：正常な嚥下, 右：誤嚥.

嚥下障害の概要を図 1.1 に示す。嚥下障害は症候名であり、広義では嚥下第 1～3 期における障害の総称であるが、一般的には嚥下第 2 期における障害にのみ用いることが多い。咽頭期における障害で生じる症状のうち最も重大なものは誤嚥であり、窒息や肺炎などの呼吸器合併症を発症し、重症化する可能性がある。嚥下障害は高齢者に多く認められ、在宅高齢者の約 3 分の 1、老年科患者のほぼ半数、介護施設入居者の半数以上に影響を及ぼしている [2]。嚥下障害発症の主な原因は脳血管疾患や神経変性疾患であるが、近年では全身性のサルコペニアも 1 つの原因と考えられている。入院患者のサルコペニアの有病率は高いことから、サルコペニアや低栄養を併存した嚥下障害患者の増加が推測される [3]。嚥下障害と栄養状態には双方向的な関連性が存在し、栄養不良も嚥下障害のリスクを高め、相互に悪循環を形成することが臨床的に重要である。これらの影響は全身状態や免疫機能にも波及し、健康予後や QOL を左右する要因となる。このように、嚥下障害の原因は多様化しており、これに応じた適切な治療を行う必要がある [4]。

摂食嚥下リハビリテーションは、嚥下機能回復を目的とした治療の一種である。食品を経口摂取する直接訓練では、嚥下機能を評価するにあたり、咽頭残留を検知することが重要である [5]。医用画像を用いた嚥下機能評価法の代表例には以下の 2 種類がある。

#### ■ 嚥下内視鏡検査 (VideoEndoscopic examination of swallowing: VE)

嚥下器官の非嚥下時の状態の観察と、少量の着色水や食物を嚥下させた際の観察を行う。ベッドサイドや在宅でも施行可能であるが、嚥下の瞬間が確認できることや、患者の苦痛等の課題がある。

#### ■ 嚥下造影検査 (VideoFluoroscopic examination of swallowing: VF)

造影剤嚥下時の嚥下器官の運動や造影剤の動きを観察する検査であり、VE では評価が難しい口腔期や食道期の観察も行うことが可能である。嚥下機能検査としては最も有用性が高く、ゴールドスタンダードとして位置付けられている。しかし、レントゲン設備の必要性や X 線の侵襲性、造影剤誤嚥等の問題から日常的な使用は制限されているため、VE の結果に基づいて必要性を判断した上で実施する。

現状、医師による診断は VF に基づく定性的評価が中心であり、嚥下機能を客観的に評価するための定量的指標は未だ確立されていない。

近年、VF に代わる非侵襲的かつ定量的な嚥下機能の評価手法として、超音波検査（嚥下エコー）が注目されている。画像による嚥下機能評価法の概要を図 1.2 に示す。

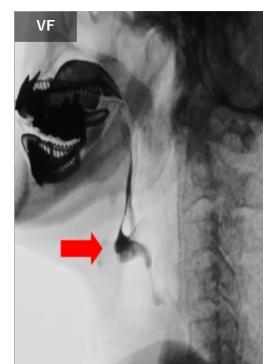


図 1.2: 画像による嚥下機能評価法の代表例。左から VE, VF, 嚥下エコーを示す。

超音波検査による嚥下機能評価法には以下3つの利点がある。

- ✓ 検査室への移動や放射線被爆を考慮する必要がないため、患者への負担が少ない。
- ✓ 造影剤を使用しないため、検査の際の食品に制限がない。
- ✓ 筋・軟部組織の描出に優れ、その形態や動きを捉えることが可能である。

しかし、超音波画像の解釈には専門医の知識を要するため、看護師や患者が日常的なリハビリテーションで活用するには困難が伴う。ゆえに、AIを用いた食塊検知の自動化が求められている。

本研究では、Gaoら[6]に着想を得て、空間的選択性を付与する注意機構[7]を導入したOne-Class Classification (OCC) 設定に従う知識逆蒸留に基づく嚥下機能評価法を提案する。本手法は、超音波検査の枠組みに導入することが可能な教師なし食塊検知モデルである。

具体的には、OCC設定に従う知識逆蒸留過程において特徴を圧縮する際に、正常特徴の効果的な保持を目的として、注目すべき特徴を強調し、かつ空間的特徴に着目するConvolutional Block Attention Module (CBAM)を組み込む。提案するOne-Class Attentive Embedding (OCAE) ブロックを採用したモデルは、正常特徴を強調しノイズ的成分を相対的に弱めることで、競合手法に比べて正常特徴への選択性が安定し、精緻化された出力を得ることが可能である。

実験では、嚥下超音波画像データセット[8]における画素単位の検知精度および領域一致度を評価することで、提案手法の有効性を示す。また、アブレーション研究や比較実験を通して、モデルの適切な構成について検討する。

## 1.2 本論文の構成

本論文は全5章で構成されている。

第1章 本研究の背景と提案手法の概要を述べ、本論文の構成を示す。

第2章 深層学習に基づく食塊検知法について述べる。

第3章 注意機構を導入した知識逆蒸留に基づく食塊検知法について述べる。

第4章 実環境を想定したデータにおいて、提案手法の評価実験について述べる。

第5章 本論文の結論を述べる。

# 第2章 深層学習に基づく食塊検知法

本章では、深層学習に基づく食塊検知法について述べる。食塊領域の推定は大きく2つのステップから構成される。すなわち、食塊が食道領域を通過するフレームのみの抽出と、食塊が通過する食道領域の推定である。本研究では、特に咽頭期における食塊の検知に焦点を当てる。

## 2.1 教師あり食塊検知法

近年の深層学習の急速な発展により、嚥下時の食道における食塊を深層ニューラルネットワークによって検知する研究が主流となっている。

Gaoら [8] は、超音波画像中の食塊特徴を活性化する注意機構に基づく教師なし食塊検知モデルを提案した。本手法は、図 2.1 に示すアテンションゲートモジュールにより予測に無関係な領域における特徴活性化を抑制し、ノイズを含む超音波画像において食塊特徴の識別能力を強化することで、視覚的に解釈可能な結果を提供する。

しかし、教師データに依存する食塊検知では食塊の粘度や患者の多様性により嚥下時の食塊形状に差が生じるため、これらの多様性を網羅的にラベル付けすることが難しく汎化能力に限界がある。

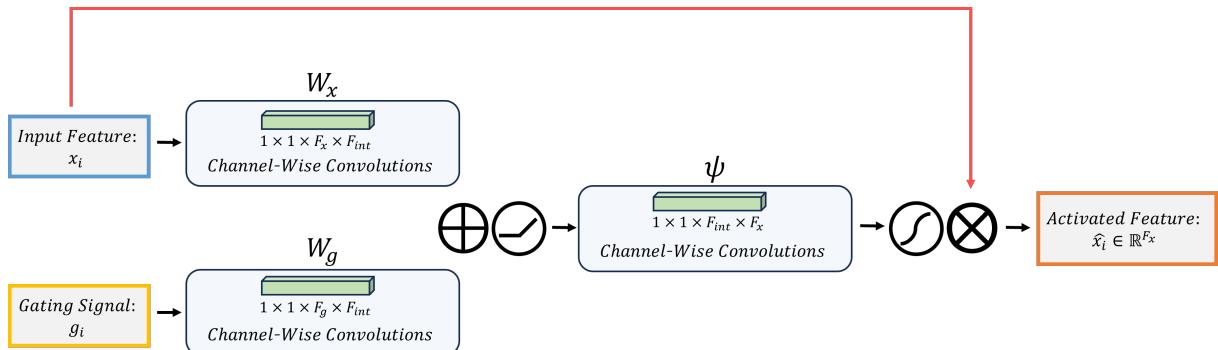


図 2.1: Attention Gate Module の概要.

## 2.2 教師なし食塊検知法

教師あり食塊検知法が抱える食塊形状の多様性に起因する問題に対処するため、近年では、嚥下動作を伴わない正常サンプルのみを用いてモデルを学習する One-Class Classification (OCC) に基づく手法 [6] が提案されている。本手法は、食塊を含まない食道状態を正常と定義し、その特徴表現分布を学習する。推論時には、学習した正常分布から逸脱する、すなわち食塊を含む特徴を異常として検知する。OCC に基づく手法は、健康状態を基準とする医師の診断フローを模倣し、データ不足に起因する制約を緩和する [9, 10]。

以下では、教師なし食塊検知法の背景にある問題設定について述べる。2.2.1 節以降では、教師なし食塊検知モデルを構成するアーキテクチャについて述べる。

### ■ 異常検知 (Anomaly Detection)

異常検知とは、データ中の異常サンプルを識別するタスクである。一般的に異常サンプルは未知であるため、学習データはラベル付けされていない正常サンプルのみで構成され、教師なし学習問題として扱われることが多い。この問題設定は以下で述べる One-Class Classification (OCC) として定式化されており、正常サンプルのみを適切に記述するモデルを学習することで異常検知を行う。

また、異常検知において様々な自己教師ありタスクの構築を試みる研究も存在する。これらのタスクには、サンプル再構築 [11]、擬似異常データ拡張、知識蒸留 [12] などが含まれる。異常検知アルゴリズムはしばしば、機械やシステムの正常稼働状態で収集されたデータに基づいて訓練され、監視に用いられる。その他の応用分野としては、サイバーセキュリティにおける侵入検知、不正検知、医療診断などがある [13]。

### ■ One-Class Classification (OCC)

OCC とは、学習中に単一のクラス（正常サンプル）のみが与えられる分類設定を指す。OCC 設定に従うモデルは、与えられた入力データが正常サンプルに属するか否かを判定することを目指す。実際の教師なし異常検知では、正常サンプルの特徴表現分布のみを学習し、推論時に正常分布から逸脱する入力を異常として判定する。

OCC 設定に従う教師なし異常検知の代表的手法としては、正常サンプルを包含する超平面または超球を学習し、推論時にその外側に分布する入力を異常とみなす One-Class Support Vector Machine (OCSVM) [11] や、正常サンプルの再構成誤差を最小化するように学習し、推論時にこの誤差が大きくなる入力を異常とみなす AutoEncoder (AE) [14] などが挙げられる。本手法は、工業的欠陥検知や医用画像処理、バイオメトリクスなど、異常サンプルを網羅的に学習することが困難な分野において幅広く利用されている。

本研究で扱う医用画像処理分野において異常検知を応用する場合、無数に存在し得る病変をラベル付けし、教師あり学習として定式化するには困難が伴う。このような条件下で、OCC 設定に従うモデルは比較的収集が容易な健康状態を学習し、学習した基準からの逸脱を病変として検知することが可能である。さらに、原理的には人間が見落とすレベルの未知または潜在的な病変をも捉える可能性を持ち、早期発見が重要な疾患に対する応用が期待される。

### 2.2.1 知識蒸留 (Knowledge Distillation)

知識蒸留 [12] とは、大規模かつ高性能な教師モデルが学習した知識を小規模かつ計算効率の高い生徒モデルへ転送することで、性能を維持しつつモデルを軽量化するフレームワークである。生徒は教師と同一または類似した Neural Network (NN) を採用し、生のデータや画像を入力として受け取り、教師の特徴活性化を模倣することを目的とする。

OCC 設定に従う知識蒸留は、教師なし異常検知の困難な課題において有望な結果を示している [15]。生徒は正常サンプルのみを学習するため、推論時に異常サンプルが入力された際に教師とは大きく異なる表現を生成することが期待される。すなわち、教師-生徒 (Teacher-Student: T-S) モデル間の異常に対する表現差異は、異常検知において重要な証拠を提供する。この仮説が、異常検知における知識蒸留ベース手法の基本となっている。

しかし、この仮説は必ずしも成立せず、実際には異常サンプルにおける活性化の差異が消失し、異常検知に失敗することがある。その原因は以下の 2 点にある。

- (i) 教師と生徒のアーキテクチャが同一または類似している、すなわち判別性のないフィルタを使用していること。
- (ii) 知識転送・蒸留時の T-S モデルにおいてデータフローが同一であること。

生徒の NN をより小さくすることでこの問題に部分的に対処する研究も存在するが、浅いアーキテクチャの表現能力の弱さが、異常の正確な検知・局在化を妨げている。

### 2.2.2 知識逆蒸留 (Knowledge Reverse Distillation)

知識蒸留が抱える問題に対処するため、Deng ら [16] は T-S モデルにエンコーダ-デコーダ構造を採用し、知識を教師の深層から浅層に蒸留する「知識逆蒸留」を提案している。本手法において、教師は生のデータや画像から豊富な特徴表現を抽出することを目的とする。生徒は教師の対称的かつ反転した構造を持ち、教師の出力を入力として、教師の挙動を模倣することを目指す。

OCC 設定に従う知識逆蒸留では、推論時に生じる T-S モデル間の再構成誤差を異常として捉える。T-S モデルの対称設計により、教師と生徒の特徴次元を一致させることが可能である。さらに、反転設計は生徒が異常特徴に対して頑健な識別性能を促す。NN の浅層は色やエッジ、テクスチャといった局所情報を抽出し、より広い受容野を持つ深層は領域的・構造的情報を考慮することが可能であるため、推論時にマルチスケール特徴に基づく蒸留を用いることで、モデルは局所的かつ領域的な異常を捉える。すなわち、教師と生徒の低次・高次特徴の類似度が低いことは、それぞれ局所的・領域的異常を示す。

知識蒸留ならびに知識逆蒸留の概要を図 2.2 に示す。従来の知識蒸留では教師と生徒の両方がエンコーダ構造を採用しているのに対し、知識逆蒸留では異種アーキテクチャ（教師はエンコーダ、生徒はデコーダ）を採用している。また、従来の T-S モデルでは生のデータや画像を教師と生徒の両方に同時に入力するが、知識逆蒸留では教師が抽出したマルチスケール特徴の低次元埋め込みを生徒に入力し、生徒は教師の出力を再構成することを目指す。回帰的観点から見れば、知識逆蒸留は生徒に教師の表現を予測させることに相当する。したがって「逆」という用語は、T-S モデルの構造と、知識蒸留過程（高次表現から低次特徴）の逆方向性を意味している。

まとめると、知識逆蒸留には以下 2 つの主要な利点がある。

✓ 非類似構造

提案する T-S モデルでは、教師はダウンサンプリングフィルタ、生徒はアップサンプリングフィルタとして機能する。これにより、判別性のないフィルタ問題 (i) を回避可能である。

✓ コンパクト埋め込み

生徒に入力される低次元埋め込みは、正常パターンの再構成における情報ボトルネックとして機能する。異常特徴を正常パターン上の微細な摂動として捉えた場合、このコンパクトな埋め込みは異常特徴の生徒への伝播を防ぎ、T-S モデル間の異常に對する表現差異を増強する。

また、注目すべき点として、従来の AE ベースの手法はピクセル差を利用して異常を検知するが、知識逆蒸留では高密度な記述特徴を用いて識別を行う。画像における領域認識的な深層特徴は、ピクセルレベルの特徴よりも効果的に異常を識別可能である。

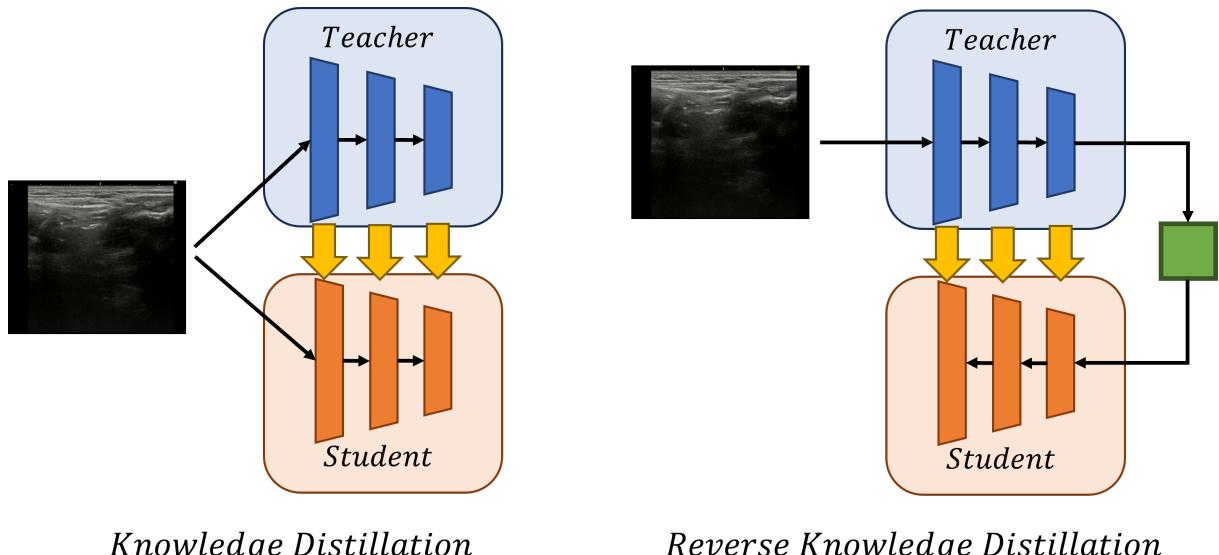


図 2.2: 知識蒸留の概要。左：知識蒸留、右：知識逆蒸留。

### 2.2.3 One-Class Bottleneck Embedding (OCBE)

知識逆蒸留における生徒は、教師の出力を入力として、教師が抽出したマルチスケール特徴を再構成することを目指す。このとき、単純に教師の最終層の出力をそのまま生徒に入力することも考えられる。しかし、この単純接続には以下2つの問題がある。

- (i) 教師は通常高い表現能力を持つが、高次元の特徴表現は情報が豊富である一方で、冗長性も高い。このような冗長特徴は、生徒が正常パターンの本質的な特徴を再構成する妨げとなる。
- (ii) 教師の最終層の特徴は主に意味的・構造的情報を表現する。知識蒸留を逆順で実施する本手法においてこれを直接生徒に入力すると、生徒側での低次特徴（色やエッジ）の再構成が難しくなる。従来のデータ再構築手法ではエンコーダとデコーダ間でショートカット接続を設けることが一般的であるが、本手法においてショートカット接続を採用すると推論時に教師の異常特徴が生徒に直接伝達され、異常情報が漏洩する。

上記の問題に包括的に対処するために、DengらはT-SモデルにOCBEモジュールを導入している。OCBEモジュールは、図2.3に示すように、教師が抽出したマルチスケール特徴を集約するMulti-Scale Feature Fusion (MFF) ブロックと、MFFブロックの出力を圧縮するOne-Class Embedding (OCE) ブロックで構成される。以下では各ブロックについて述べる。

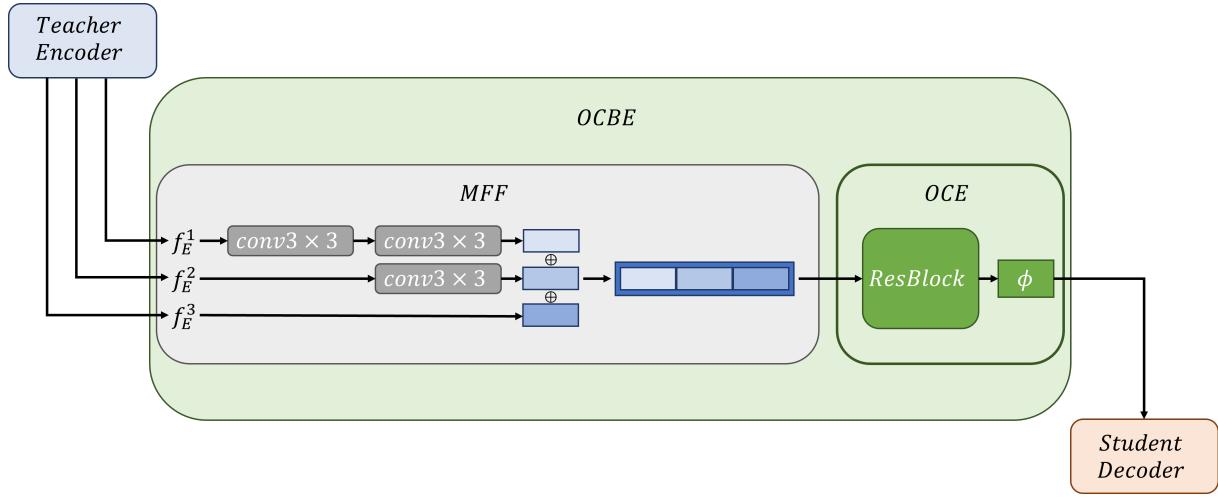


図 2.3: OCBE モジュールの概要.

## ■ Multi-Scale Feature Fusion (MFF)

前述の問題 (i) に対処するために, MFF ブロックでは OCE の前に教師から低次および高次特徴を集約し, 正常パターン再構成のための豊かな埋め込みを構築する.

特徴結合時の整合性を確保するため, 浅層の特徴は  $3 \times 3$  の畳み込み層 (stride 2) でダウンサンプリングし, バッチ正規化と ReLU 活性化を施す. その後,  $1 \times 1$  の畳み込み層 (stride 1) とバッチ正規化, ReLU を用いて, 冗長性の少ない豊富な特徴表現を得る.

## ■ One-Class Embedding (OCE)

前述の問題 (ii) に対処するために, OCE ブロックでは MFF で得られた特徴をコンパクトなボトルネックコードに圧縮し, 生徒が教師の出力を再構成するために必要な重要情報, すなわち正常特徴を効果的に保持する. これにより, 異常時に T-S モデル間で教師と生徒の表現差異がより顕在化する. Deng らは, OCE ブロックとして ResNet [17] の第 4 残差ブロック (ResBlock) を使用している.

図 2.3において, MFF ブロックの畳み込み層ならびに OCE ブロックを構成する ResBlock は学習可能であり, 生徒と共に正常サンプルにより最適化される. 実際に, 超音波画像はノイズが多いため, 教師の高次元特徴を OCBE モジュールによりコンパクトな潜在空間に射影することで, 正常パターンである食塊を含まない食道状態の特徴表現を効果的に保持することが可能である.

## 2.2.4 知識蒸留損失 (KD Loss)

知識逆蒸留では、高次元・低次元双方での特徴関係を適切に捉える知識蒸留損失 [18] により OCBE モジュールと生徒デコーダが共同で最適化される。

形式的には、入力データ  $I$  から OCE 空間への射影を  $\phi$  とすると、教師エンコーダと生徒デコーダの対応する活性化は次のようになる：

$$\begin{cases} f_E^k = E^k(I), \\ f_D^k = D^k(\phi). \end{cases} \quad (2.1)$$

$E^k$  と  $D^k$  は、それぞれ  $k$  番目のエンコード・デコードブロックを表す。また、特徴テンソルは  $f_E^k, f_D^k \in \mathbb{R}^{C_k \times H_k \times W_k}$  であり、 $C_k, H_k, W_k$  はそれぞれチャネル数、高さ、幅を表す。

知識蒸留損失では、 $f_E^k$  と  $f_D^k$  に対してチャネル軸に沿ったベクトル単位のコサイン類似度損失を計算することで、2次元異常度マップ  $M^k \in \mathbb{R}^{H_k \times W_k}$  を得る：

$$M^k(h, w) = 1 - \frac{(f_E^k(h, w))^T \cdot f_D^k(h, w)}{\|f_E^k(h, w)\| \|f_D^k(h, w)\|}. \quad (2.2)$$

$M^k(h, w)$  において大きな値は、位置  $(h, w)$  において異常の可能性が高いことを示す。

マルチスケール特徴に基づく蒸留を考慮したスカラー損失関数は、異常度を各スケール全位置で平均し、その後全スケールで加算することで定義される：

$$\mathcal{L}_{KD} = \sum_{k=1}^K \left\{ \frac{1}{H_k W_k} \sum_{h=1}^{H_k} \sum_{w=1}^{W_k} M^k(h, w) \right\}. \quad (2.3)$$

$K$  は考慮する特徴階層数を表す。

推論時には、式 (2.2) に従い各層ごとにピクセルレベルの再構成誤差を計算し、 $K$  枚の異常度マップを得る。最終的な画像全体の異常度マップ  $S_{AL}$  は、 $M^k$  をバイリニア補間  $\Psi$  により画像サイズにアップサンプリングしたのち、ピクセルレベルで全スケールを加算することで得る：

$$S_{AL} = \sum_{i=1}^L \Psi(M^i). \quad (2.4)$$

得られたスコアマップ  $S_{AL}$  はガウシアンフィルタで平滑化し、ノイズを除去する。

画像単位の判定においては、異常度マップ全体の平均値を用いると小さな異常領域を持つサンプルが不利になる。そのため、最も強く反応したピクセルのスコアをサンプルレベルの異常スコアと定義する。正常サンプルの場合、スコアマップ上に顕著な反応は存在しないと仮定する。

# 第3章 注意機構を導入した知識逆蒸留に基づく食塊検知法の提案

本章では、提案手法として教師なし食塊検知モデルに組み込む注意機構のアーキテクチャとその導入方法について述べる。

具体的には、T-S モデルにおける OCBE モジュールが OCE ブロックにて情報を圧縮する際に、注目すべき特徴を強調するチャネルアテンション、ならびに空間的特徴に着目する空間アテンションを併用する注意機構である CBAM [7] を組み込み、OCAE ブロックとして定式化する。

## 3.1 注意機構 (Attention Mechanism)

注意機構とは、入力特徴の中からタスク遂行に有用な情報を選択的に強調し、相対的に重要度の低い情報を抑制する計算機構であり、query, key, value に基づき各特徴量の重要度を算出し、重み付き和として集約する。これにより、モデルは入力全体を一様に処理するのではなく、重要な特徴に重点を置いた表現を生成する。元来、注意機構は系列変換問題において長距離依存関係を効率的に捉える目的で導入され、その有効性が示されてきた [19]。近年では Convolutional Neural Network (CNN) にも拡張され、特徴の重要度を動的に調整するアテンションモジュールとして発展している。特に大規模分類タスクにおいて、CNN の性能向上のために注意機構を組み込む研究は数多く存在する。Wang ら [20] は、エンコーダ-デコーダ型のアテンションモジュールを用いる Residual Attention Network を提案した。本手法では 3 次元アテンションマップを直接計算し特徴マップを精緻化することで、性能を高めると同時にノイズの多い入力に対しても頑健性を持たせている。また、Hu ら [21] はチャネル間の関係性を活用するコンパクトなモジュールを提案している。本手法の Squeeze-and-Excitation モジュールでは、グローバル平均プーリングした特徴を用いてチャネルごとのアテンションを計算する。

本研究では、食道と食塊の空間的位置関係を捉えることが重要であると考え、Woo ら [7] が提案する CBAM を採用する。Woo らは、3 次元特徴マップに対してアテンションを分離して生成することで、Wang らの手法と比較して計算量およびパラメータ数が削減されるために、既存の CNN に容易に導入可能であると主張している。また、空間情報の集約には一般的に平均プーリングが用いられてきたが、Woo らはより顕著な特徴に注目する最大プーリングも併用することを提案している。さらに、効率的なアーキテクチャに基づき、チャネルアテンションと同時に「どこに」注目すべきかを決定する上で重要な空間アテンションを併用することが、チャネルアテンション単独の使用よりも優れていることを検証している。4.3.1 節で実施するアブレーション研究では、提案手法における各モジュールの有効性について確認する。以下では、CBAM のアーキテクチャと提案手法における CBAM の導入方法について述べる。

## 3.2 Convolutional Block Attention Module (CBAM)

CBAM の概要を図 3.1 に示す。本モジュールは、チャネルアテンションモジュールと空間アテンションモジュールの逐次的配置からなる。

形式的には、中間特徴マップ  $\mathbf{F} \in \mathbb{R}^{C \times H \times W}$  に対して、1 次元のチャネルアテンションマップ  $\mathbf{M}_c(\mathbf{F}) \in \mathbb{R}^{C \times 1 \times 1}$  と 2 次元の空間アテンションマップ  $\mathbf{M}_s(\mathbf{F}') \in \mathbb{R}^{1 \times H \times W}$  を順次推定する。すなわち、最終的なアテンションの出力  $\mathbf{F}''$  は次式で計算される：

$$\begin{cases} \mathbf{F}' = \mathbf{M}_c(\mathbf{F}) \otimes \mathbf{F}, \\ \mathbf{F}'' = \mathbf{M}_s(\mathbf{F}') \otimes \mathbf{F}'. \end{cases} \quad (3.1)$$

$\otimes$  は要素ごとの積を表し、チャネルアテンション値は空間次元、空間アテンション値はチャネル次元に沿ってそれぞれブロードキャストされる。

チャネルアテンションと空間アテンションの 2 つのモジュールは、互いに補完的なアテンションを計算することで、それぞれ次元および空間で注目すべき特徴に焦点を当てる。また、各アテンションは Woo らに従い、チャネルアテンション、空間アテンションの順に逐次配列する。4.3.3 節で実施する比較実験では、提案手法における CBAM の適切な構成について検討する。

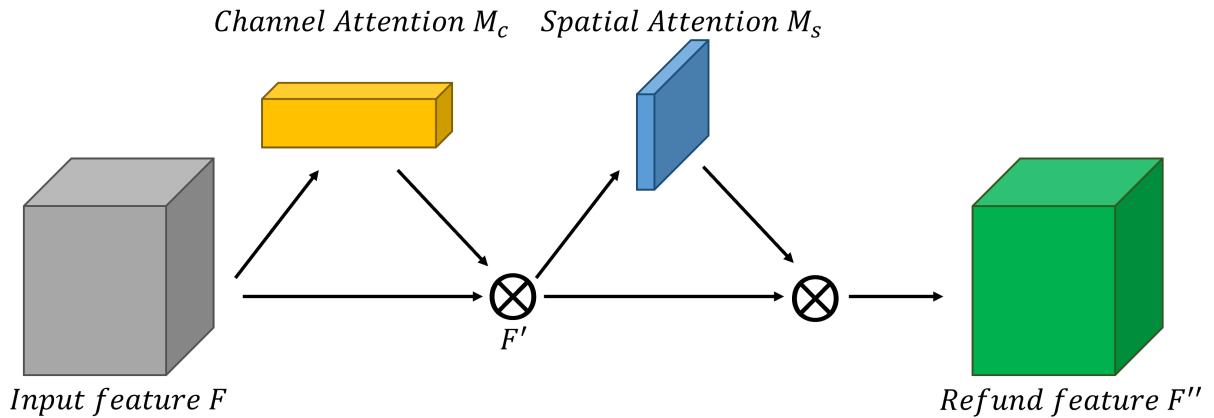


図 3.1: CBAM の概要。

### 3.2.1 Channel Attention Module

チャネルアテンションモジュールの概要を図 3.2 に示す。チャネルアテンションマップは、特徴のチャネル間関係を利用して生成する。特徴マップの各チャネルは一種の特徴検出器とみなせるため、チャネルアテンション  $M_c$  は入力画像に対して注目すべき特徴表現に焦点を当てる [22]。

形式的には、特徴マップに平均プーリングと最大プーリングを適用し、 $\mathbf{F}_{\text{avg}}^c, \mathbf{F}_{\text{max}}^c \in \mathbb{R}^{C \times 1 \times 1}$  を得る。これらはそれぞれ平均プーリングされた特徴および最大プーリングされた特徴を表す。 $\mathbf{F}_{\text{avg}}^c, \mathbf{F}_{\text{max}}^c$  は 1 つの隠れ層  $\mathbb{R}^{C/r \times 1 \times 1}$  を持つ MLP に入力される。その後、MLP を通過した出力特徴ベクトルは要素ごとの和で統合され、チャネルアテンションマップ  $M_c(\mathbf{F}) \in \mathbb{R}^{C \times 1 \times 1}$  が生成される。すなわち、チャネルアテンションマップは次式で計算される：

$$\begin{aligned} M_c(\mathbf{F}) &= \sigma(MLP(\text{AvgPool}(\mathbf{F})) + MLP(\text{MaxPool}(\mathbf{F}))) \\ &= \sigma(\mathbf{W}_1(\mathbf{W}_0(\mathbf{F}_{\text{avg}}^c)) + \mathbf{W}_1(\mathbf{W}_0(\mathbf{F}_{\text{max}}^c))) \end{aligned} \quad (3.2)$$

$\sigma$  はシグモイド関数を表す。また、 $\mathbf{W}_0 \in \mathbb{R}^{C/r \times C}$ ,  $\mathbf{W}_1 \in \mathbb{R}^{C \times C/r}$  はそれぞれ MLP における特徴変換行列、 $r$  は MLP における次元縮小率を表す。

本研究では、Woo らに従い  $r = 16$  を用いる。4.3.4 節で実施する比較実験では、提案手法におけるチャネルアテンションモジュールの適切な次元縮小率について検討する。

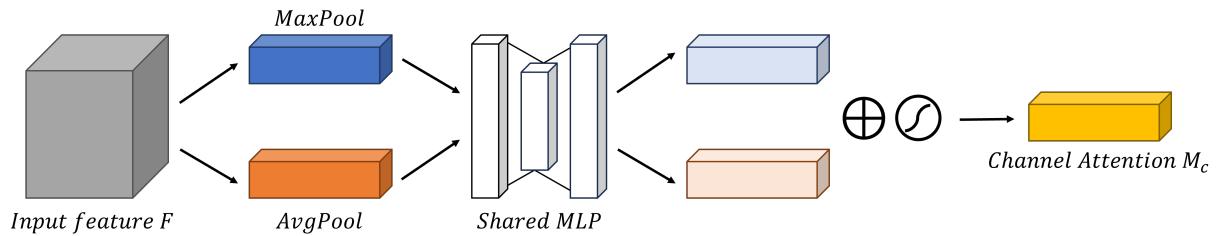


図 3.2: Channel Attention Module の概要.

### 3.2.2 Spatial Attention Module

空間アテンションモジュールの概要を図 3.3 に示す。空間アテンションマップは、特徴の空間的関係を利用して生成する。空間アテンション  $M_s$  は、チャネルアテンションで失われる空間情報を補完する。

形式的には、特徴マップにチャネル方向の平均プーリングおよび最大プーリングを適用し、2次元マップ  $\mathbf{F}'^s_{\text{avg}}, \mathbf{F}'^s_{\text{max}} \in \mathbb{R}^{1 \times H \times W}$  を得る。チャネル軸に沿ったプーリング操作は、有益な領域を強調するのに効果的であることが示されている [23]。 $\mathbf{F}'^s_{\text{avg}}, \mathbf{F}'^s_{\text{max}}$  の連結後、畳み込み層を通して2次元の空間アテンションマップ  $M_s(\mathbf{F}') \in \mathbb{R}^{1 \times H \times W}$  が生成される。すなわち、空間アテンションマップは次式で計算される：

$$\begin{aligned} M_s(\mathbf{F}') &= \sigma\left(f^{7 \times 7}([AvgPool(\mathbf{F}'); MaxPool(\mathbf{F}')])\right) \\ &= \sigma\left(f^{7 \times 7}([\mathbf{F}'^s_{\text{avg}}; \mathbf{F}'^s_{\text{max}}])\right) \end{aligned} \quad (3.3)$$

$\sigma$  はシグモイド関数、 $f^{7 \times 7}$  はカーネルサイズ  $7 \times 7$  の畳み込み演算を表す。

本研究では、Woo らに従い  $k = 7$  を用いる。4.3.4 節で実施する比較実験では、提案手法における空間アテンションモジュールの適切なカーネルサイズについて検討する。

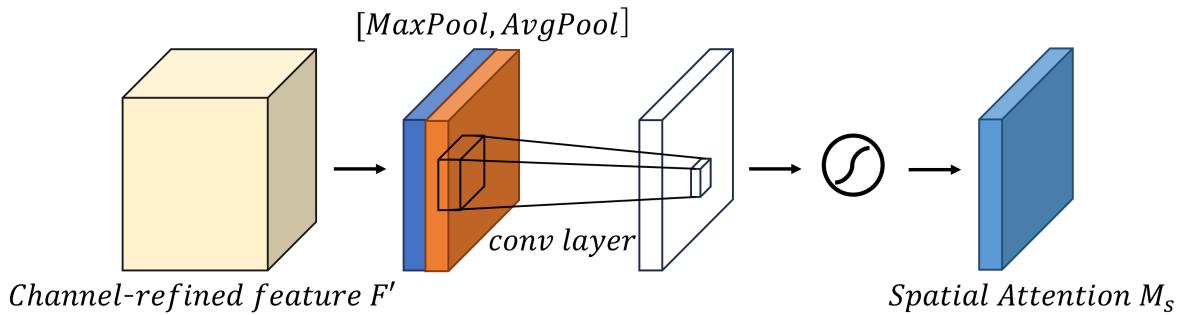


図 3.3: Spatial Attention Module の概要。

### 3.3 OCAE ブロックを用いた知識逆蒸留に基づく食塊検知法

提案手法の概要を図 3.4 に示す。提案手法は、教師エンコーダ  $E$ 、MFF ブロックと提案する OCAE ブロック  $\phi$  からなる学習可能な OCBE モジュール、生徒デコーダ  $D$  で構成される知識逆蒸留に基づく教師なし食塊検知モデルである。

教師エンコーダのパラメータは学習時に固定され、特徴抽出能力の低下を防ぐ。CBAM を組み込んだ OCBE モジュールは、MFF ブロックにて教師が抽出したマルチスケール特徴を集約し、OCAE ブロックにて特徴  $x$  を低次元空間に埋め込むとともに、チャネルおよび空間方向のアテンションにより正常特徴を効果的に保持した潜在表現  $z = \phi(x)$  を生成する。生徒デコーダは  $z$  を入力として教師が抽出したマルチスケール特徴を再構成するように学習を行う。OCBE モジュールおよび生徒デコーダは正常サンプルの再構成に最適化されているため、推論時に異常サンプルが入力された場合、教師エンコーダが抽出した特徴と生徒デコーダによる再構成との誤差を異常として検知することが可能である。

提案手法のモデルは、エンコーダ、OCBE モジュール、デコーダのバックボーンネットワークとして、過度な深層化による最適化困難や計算非効率を避けるため、ResNet [17]において残差ブロックの数を増やす代わりにチャネル数を拡張した残差ネットワークである Wide Residual Network (WRN) [24] を採用する。

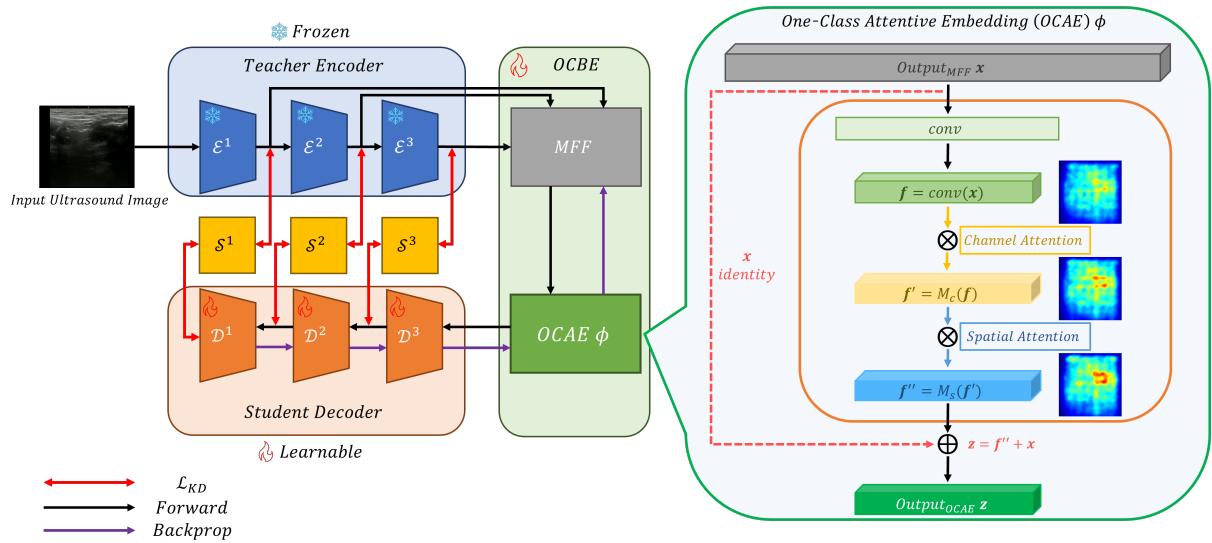


図 3.4: 提案手法の概要.

### 3.3.1 Residual Network (ResNet)

CNNは、積層数（深さ）によって特徴のレベルを豊かにすることが可能である。しかし、ネットワークの深さを単純に増加させると勾配消失や性能劣化といった問題が生じる。ResNetは、深層残差学習フレームワークの導入により、非常に深いネットワークにおける最適化を可能とする。

形式的には、いくつかの積層によって近似されるべき基底写像を  $H(x)$  とし、 $x$  をそれらの層への最初の入力とする。複数の非線形層が複雑な関数を漸近的に近似可能であると仮定し、積層に基底写像ではなく残差関数を漸近的に近似させる。すなわち、残差関数は次式で計算される：

$$F(x) := H(x) - x \quad (3.4)$$

このとき、元の関数は  $F(x) + x$  として表される。追加された層が恒等写像として構成可能であれば、より深いモデルの誤差はそれより浅いモデルの学習誤差以下であるはずである。恒等写像  $H(x)$  が最適であれば、残差関数  $F(x)$  は単にゼロに収束すればよい。ResNetは、いくつかの積層ごとに残差学習を導入する。基本構成ブロックの概要を図 3.5 に示す。

形式的には、次のように定義される構成ブロックを考える：

$$y = F(x, \{W_i\}) + x. \quad (3.5)$$

$x, y$  は、それぞれ対象とする層の入力ベクトルおよび出力ベクトルを表す。関数  $F(x, \{W_i\})$  は、学習される残差写像を表す。

図に示した2層からなる例では  $y = W_2\sigma(W_1x)$  であり、 $\sigma$  は ReLU を表す。簡単のため、バイアス項は省略している。 $F + x$  の演算はショートカット接続および要素ごとの加算によって実行され、直後に2つ目の非線形性を適用し  $\sigma(y)$  となる。式におけるショートカット接続は、追加のパラメータも計算量の増加も導入しない [17]。

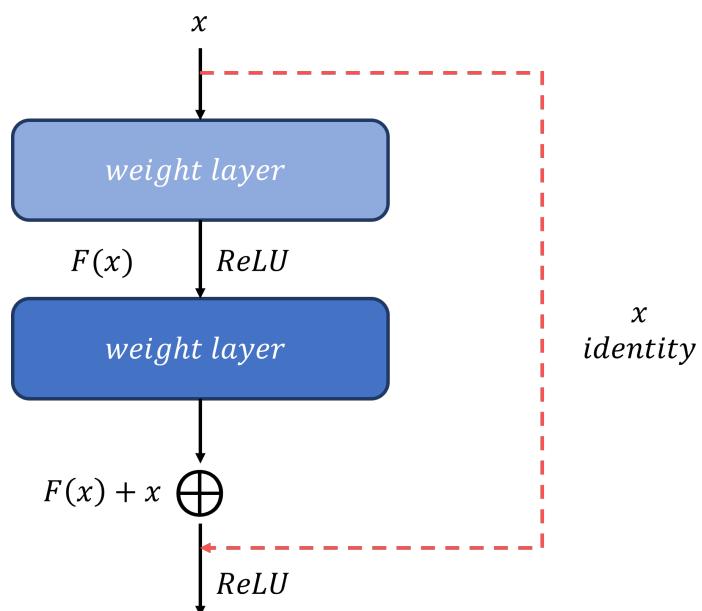


図 3.5: ResNet の概要。

### 3.3.2 Wide Residual Network (WRN)

ResNetは数千層規模までスケール可能であり、かつ性能が向上し続けることが示されている。しかし、精度がわずかに数分の1%向上するごとに必要な層数はほぼ倍増し、その結果、非常に深い残差ネットワークの学習では特徴の再利用が次第に低下するという問題が生じ、学習が非常に遅くなる。

これらの問題に対処するため、Zagoruykoら[24]は残差ネットワークの深さを減らし、チャネル数（幅）を増やす新しいアーキテクチャを提案した。本手法のWRNは、従来の細く非常に深い残差ネットワークよりもはるかに優れた性能を示す。

本研究ではGaoら[6]に従い、バックボーンネットワークとしてWRN-50-2を採用する。WRN-50-2は、図3.6に示すように、ResNet-50を基礎とした深層残差ネットワークであり、ボトルネック型残差ブロックを用いた全50層の学習可能な重み付き層から構成される。ネットワークは、初段の $7 \times 7$ 畳み込み層および最大プーリング層に続き、4つの残差ブロックを順に積層する。各ステージに配置される残差ブロック数はそれぞれ3, 4, 6, 3個であり、合計16個の残差ブロックが直列に配置される。各残差ブロックは、 $1 \times 1$ 畳み込み、 $3 \times 3$ 畳み込み、 $1 \times 1$ 畳み込みから構成され、それぞれチャネル次元の削減、空間的特徴抽出、チャネル次元の再拡張を担う。残差ブロックの出力はショートカット接続を介して入力と要素ごとに加算され、残差学習を実現する。これらの残差ブロックに含まれる主経路上の畳み込み層は合計48層であり、初段の畳み込み層および最終の全結合層を含めることで、ネットワーク全体の深さは50層となる。

標準的なResNet-50と比較して、WRN-50-2は各残差ブロック内部の畳み込み層におけるチャネル数を一様に2倍に拡張している。一方、残差ブロック数、ステージ構成、ネットワークの深さはResNet-50と同一に保たれている。この設計により、極端な深層化を行うことなく、各層における表現能力を向上させることが可能となる。ステージ間で特徴マップの空間解像度やチャネル次元が変化する場合には、ショートカット経路に $1 \times 1$ 畳み込みによる線形射影を導入して次元整合を行い、次元が一致する場合には恒等写像によるショートカット接続を用いる。すなわち、各残差ブロックの1層目の構築時のみstride 2で次元変換を行う。

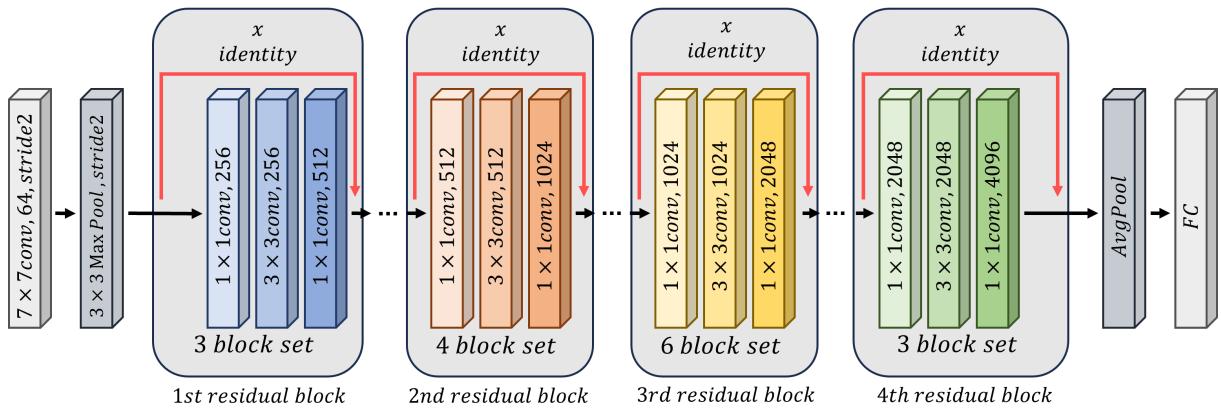


図3.6: WRN-50-2の概要。

### 3.3.3 残差ブロックに対する CBAM の導入

提案手法のモデルでは、WRN-50-2の残差ブロックを各ステージごとに分割し、特徴抽出基盤として利用する。

具体的には、第1～第3残差ブロックをエンコーダとして採用し、入力画像から段階的に高次の特徴表現を抽出する。第4残差ブロックはOCBEモジュールに割り当てられ、エンコーダにより抽出された特徴表現の圧縮および異常検知に効果的な正常特徴の保持を行い、異常特徴の伝播を抑制する。一方、デコーダはエンコーダに対して反転かつ対称的に設計される。すなわち、エンコーダにおける残差ブロックの積層順およびチャネル構成を逆転させることでデコーダを構成し、各段において空間解像度およびチャネル次元を段階的に復元する。この設計により、エンコーダで抽出されたマルチスケール特徴を効率的に再構成することが可能となり、知識逆蒸留におけるT-Sモデル間の再構成誤差に基づく食塊検知を実現する。

CBAMは残差ブロックに対して導入可能であり、その組み込み位置には複数候補があるが、本研究ではOCBEモジュールのOCEブロックを構成する残差ブロックを対象とし、OCAEブロックとして定式化する。具体的には、Wooらに従い、図3.7に示すように残差接続の直前に組み込む。4章4.3.2節ならびに4.3.4節で実施する比較実験では、CBAMを組み込む適切な位置およびWRNの適切なスケールについて検討する。

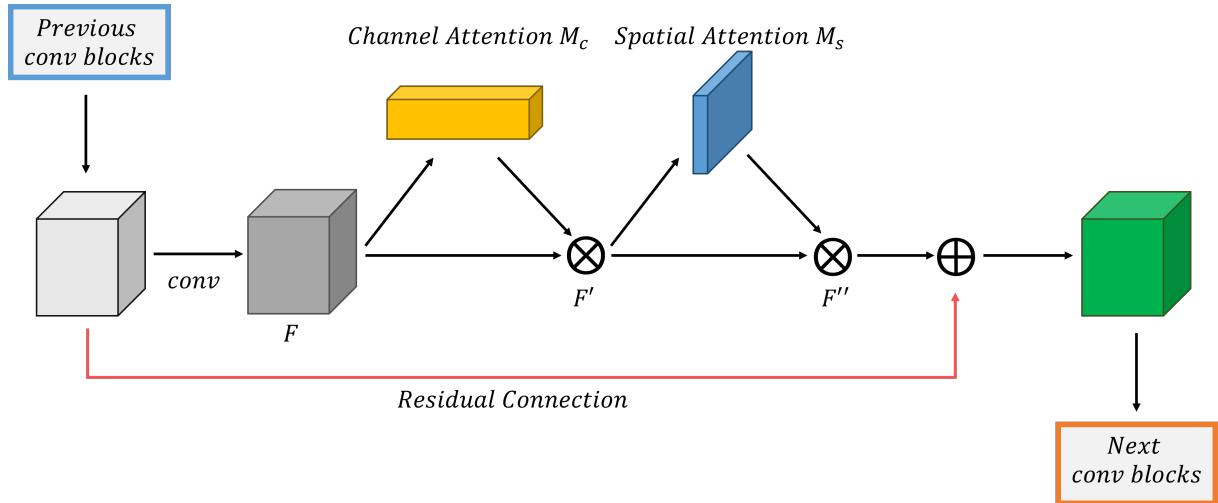


図3.7: 残差ブロックに対するCBAMの導入。

# 第4章 評価実験

本章では、嚥下超音波画像データセット [8]において画素単位の検知精度および領域一致度を評価する。4.1.2節ではGaoら [6]の手法を自環境で実装し、空間的な選択性を与える注意機構を導入した提案手法の優位性を示す。また、4.3.1節ではアブレーション研究を実施し、提案手法における各モジュールの有効性を検証する。さらに、提案手法における適切なモデル構成を実現すべく、4.3.2節ではCBAMの導入位置、4.3.2節ではCBAMの構成、4.3.4節ではハイパーパラメータおよびバックボーンネットワークについて検討する。

## 4.1 実験設定

### 4.1.1 嚥下超音波画像データセット

本研究で使用する嚥下超音波画像データセットは、適切な評価実験のためにGaoらが作成したデータセットを一部整形済みであり、7種類の食塊（Bread, Cracker, Jelly, Pudding, Soda, Yogurt, Yokan）を含む食道および空の食道における嚥下反射を捉えた超音波画像2395枚で構成される。このうち652枚を学習、1743枚をテストとして用いる。

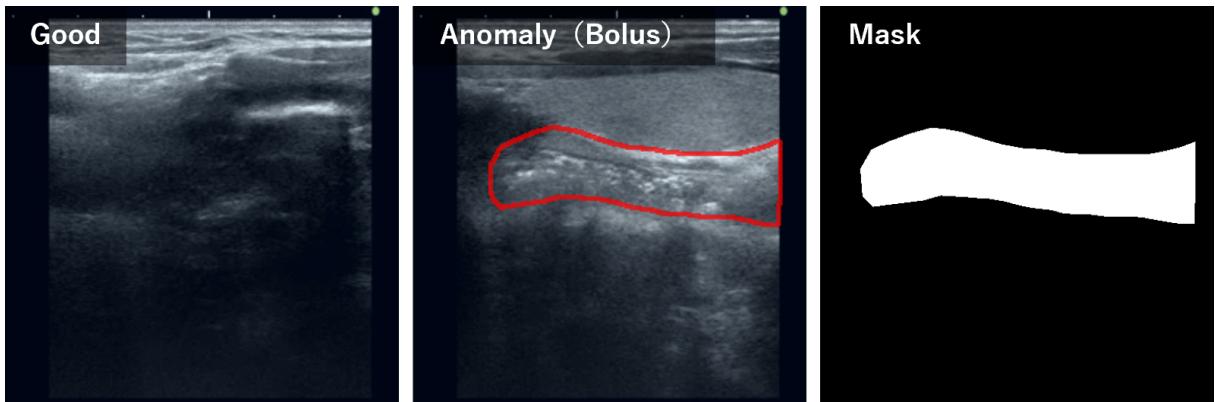


図 4.1: 嚥下超音波画像データセットの概要。左から正常画像、異常画像、マスク画像を示す。

#### 4.1.2 評価指標

評価指標には、画素単位の検知精度／領域一致度を測る Pixel AUROC ／ AUPRO を使用する。以下では、各評価指標について述べる。

##### ■ Area Under the Receiver Operating Characteristic curve (AUROC)

AUROC は、モデルの判別性能を評価する定量的指標である。ROC (Receiver Operating Characteristic) 曲線は、閾値を連続的に変化させた際に得られる真陽性率 (True Positive Rate: TPR) と偽陽性率 (False Positive Rate: FPR) を ( $FPR, TPR$ ) としてプロットした曲線であり、2 値分類器の性能を包括的に評価するための代表的手法である。

形式的には、ROC 曲線を FPR について積分した値、すなわち曲線下の面積であり、0 から 1 の範囲を取ることで分類器の全体的な識別能力を单一のスカラー値として表現する。AUROC が 0.5 の場合はランダムな判別と同等であり、0.8 以上であれば十分に良好な識別性能とされる [25]。また、AUROC はクラス分布の影響を受けにくいという特性を持つため、不均衡データを含む設定においても安定した性能比較が可能である。この特性から、本研究では異なる条件間におけるモデル性能を公平に比較する指標として AUROC を採用する。

##### ■ Area Under the Per-Region Overlap curve (AUPRO)

AUPRO は、異常領域のセグメンテーション性能を評価する定量的指標である。PRO (Per-Region Overlap) は、正解異常領域に対する予測異常領域の被覆率を平均的に評価する指標であり、次式で計算される：

$$PRO(T, P) = \frac{1}{K} \sum_{k=1}^K \frac{|P \cap T_k|}{|T_k|} \quad (4.1)$$

$K$  は正解異常領域の数、 $T_k$  は  $k$  番目の正解異常領域、 $P$  は予測された異常画素の集合を示す。

PRO 曲線は、閾値を連続的に変化させた際に得られる PRO と FPR を ( $FPR, PRO$ ) としてプロットした曲線であり、異常検知における局在化性能の評価に適している。

形式的には、PRO 曲線を FPR について積分した値、すなわち曲線下の面積である。Lin ら [26] は、AUPRO は AUROC と同様に閾値非依存な指標である一方、異常領域単位での被覆率を直接評価可能である点に特徴があり、画素数の偏りが大きい異常セグメンテーション問題において有効であると述べている。本研究では Lin らに従い、実環境を考慮した低 FPR (< 0.3) における AUPRO を採用する。

## 4.2 実験結果

本節では、競合手法 [6] を自環境で実装し、空間的な選択性を付与する注意機構を導入した提案手法の優位性を示す。表 4.1 に定量的評価、図 4.2 に定性的評価を示す。また、提案手法におけるカテゴリ別の定性的評価を図 4.3 に示す。各評価指標の列において、評価指標ごとに最も良い結果は太字で示されている。

提案手法は、多くのカテゴリで競合手法を上回ることが確認できる。OCAE ブロックは、チャネルおよび空間方向の注意を与えることで、図 4.2 に示すように正常特徴を強調しノイズ的成分を相対的に弱める。そのため、OCE ブロックにて単純な畳み込み層を用いて特徴圧縮をする競合手法に比べて、正常特徴への選択性が安定し精緻化された出力を得ることが可能である。一方、提案手法は、推論時にノイズや異常特徴が CBAM に入力されると不適切な強調を生じる可能性がある。特に、食塊領域が小さくかつ輝度差が低い画像では、図 4.4 に示すように食塊領域を大域的に捉えてしまい、予測結果に悪影響を及ぼすことがある。Yogurt, Yikan の 2 種類では本研究で使用したデータセットの特性上このようなパターンが多く、精度が僅かに低下したと解釈できる。しかし、実際にはこの懸念よりも正常特徴に対する強調効果が優勢となり、結果的に多くのカテゴリで検知精度の向上に寄与していると考えられる。

表 4.1: 嘉下超音波画像データセット [8] における検知精度の定量的評価  
(上段: AUROC (%) ↑, 下段: AUPRO (%) ↑).

Category	Bread	Cracker	Jelly	Pudding	Soda	Yogurt	Yikan	Avg
AD [6]	83.6	81.7	82.6	83.0	78.3	<b>85.1</b>	<b>84.7</b>	82.7
	62.2	53.9	57.3	56.0	41.7	<b>60.6</b>	<b>59.2</b>	55.8
Ours	<b>86.2</b>	<b>84.8</b>	<b>85.3</b>	<b>85.4</b>	<b>81.2</b>	84.8	84.5	<b>84.6</b>
	<b>64.4</b>	<b>55.4</b>	<b>59.9</b>	<b>59.5</b>	<b>44.7</b>	59.4	59.0	<b>57.5</b>

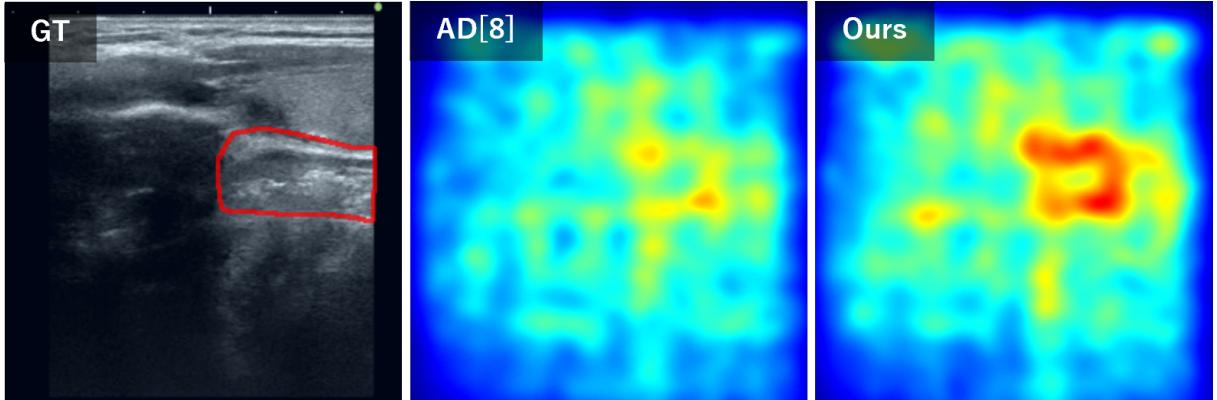


図 4.2: 競合手法 [6] ならびに提案手法の定性的評価。左から GT, 競合手法, 提案手法の異常度マップを示す。

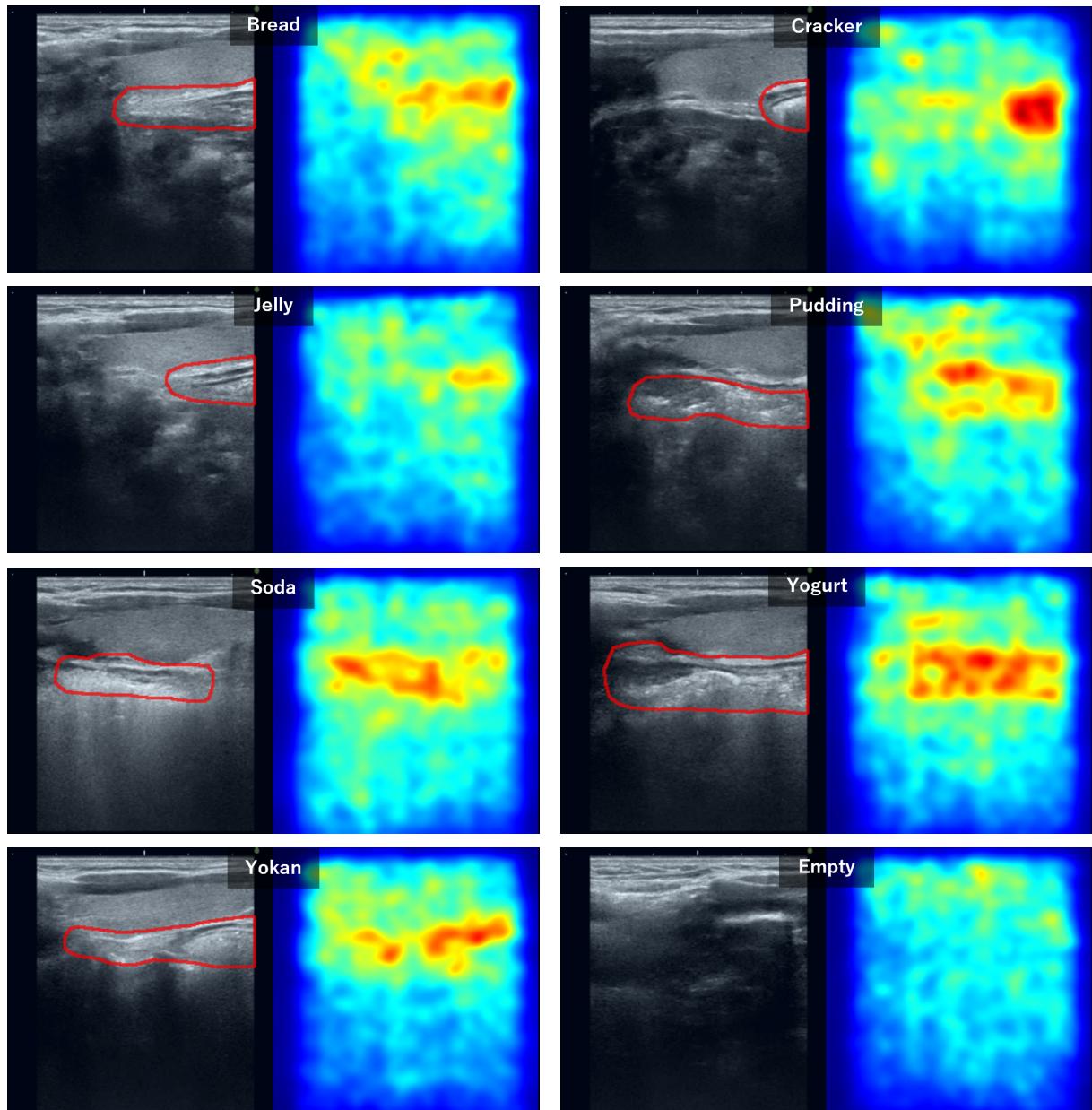


図 4.3: 7種類の食塊および空の食道における提案手法の定性的評価. 左: GT, 右: 異常度マップ.

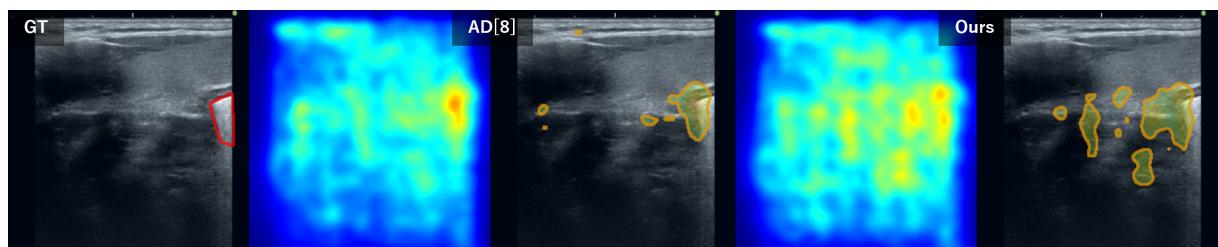


図 4.4: 競合手法ならびに提案手法の定性的評価. 左から GT, 競合手法, 提案手法の異常度マップ, 予測結果を示す.

## 4.3 分析

### 4.3.1 アブレーション研究

本節では、提案手法における各モジュールの有効性を検証すべく、嚥下超音波画像データセットにおけるアブレーション研究を実施する。表 4.2 に定量的評価、図 4.5 に定性的評価を示す。

提案手法から CBAM を除いた場合、2つの評価指標において検知精度の低下が確認できる。特に、空間アテンションモジュールを除いた場合の AUPRO の低下から、チャネル方向のみの強調では空間的分布を十分に捉えきれないことが示唆される。比較的低下率は小さいものの、チャネルアテンションモジュールを除いた場合も同様の傾向が確認できる。

また、OCBE モジュールと CBAM を除いた場合、2つの評価指標において大きく検知精度が低下することが確認できる。これは、情報が豊富な高次元特徴を抽出可能な教師エンコーダは潜在特徴の冗長性が高いために、生徒デコーダが正常パターンの本質的な特徴の再構成を妨げることが要因と考えられる。実際に図 4.5 から、知識逆蒸留過程において教師の高次表現を生徒に直接入力すると、生徒側での低次特徴の再構成が困難となることが確認できる。したがって、提案手法は空間およびチャネルの双方向から食塊検知に有効である特徴を効果的に強調し、検知精度向上に寄与すると解釈できる。

表 4.2: アブレーション研究における検知精度の定量的評価  
(上段: AUROC (%) ↑, 下段: AUPRO (%) ↑).

Category	Bread	Cracker	Jelly	Pudding	Soda	Yogurt	Yokan	Avg
w/o OCBE & CBAM	80.9	78.2	80.7	80.0	74.2	80.9	79.1	79.1
	52.7	42.2	48.4	45.2	31.7	47.3	47.2	45.0
w/o CBAM	83.6	81.7	82.6	83.0	78.3	85.1	<b>84.7</b>	82.7
	62.2	53.9	57.3	56.0	41.7	<b>60.6</b>	<b>59.2</b>	55.8
w/o Spatial	84.5	82.4	83.7	84.2	80.3	<b>85.5</b>	84.1	83.5
	62.0	52.3	56.9	56.8	43.1	60.1	59.0	55.7
w/o Channel	<b>86.2</b>	84.2	<b>85.3</b>	85.1	<b>81.7</b>	85.2	84.3	<b>84.6</b>
	64.1	53.8	59.5	58.5	<b>45.7</b>	59.7	58.6	57.1
Ours	<b>86.2</b>	<b>84.8</b>	<b>85.3</b>	<b>85.4</b>	81.2	84.8	84.5	<b>84.6</b>
	<b>64.4</b>	<b>55.4</b>	<b>59.9</b>	<b>59.5</b>	44.7	59.4	59.0	<b>57.5</b>

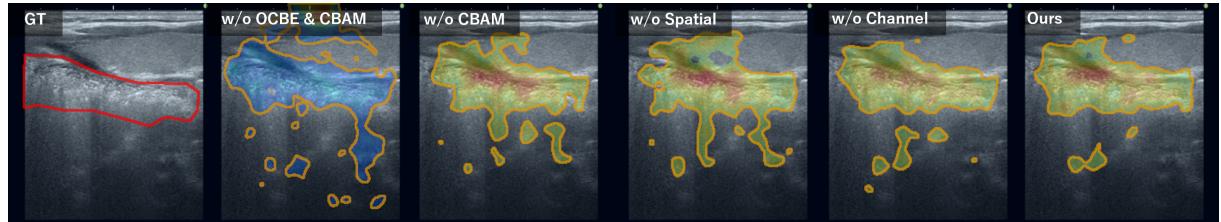


図 4.5: アブレーション研究における定性的評価。左から GT、すべてのモジュールの除去、CBAM の除去、Spatial Attention のみ除去、Channel Attention のみ除去、提案手法の予測結果を示す。

### 4.3.2 CBAM の導入位置に関する検討

本節では、提案手法における適切なモデル構成を実現すべく、CBAM の導入位置について検討する。提案手法のモデルはエンコーダ、OCBE モジュール、デコーダが全て WRN [24] で構成されている。そこで、本モデルにおいて異なるモジュールに CBAM を導入した場合の検知精度を比較する。表 4.3 に定量的評価、図 4.6 に定性的評価を示す。

エンコーダの全層に CBAM を組み込んだ場合、モデルは全カテゴリにおいて最も低い検知精度を示した。これは、エンコーダ側で CBAM により誤強調されたノイズが OCBE モジュールおよびデコーダに伝播することが要因と考えられる。仮にノイズが OCBE モジュールにより抑制されたとしても、これは T-S モデル間の誤差となるため異常として検知されてしまう。また、エンコーダとデコーダの全層に CBAM を組み込んだ場合、ノイズの不適切な強調による誤差が緩和されるために検知精度の向上が確認できる。このとき、食塊領域が明瞭な画像が多く含まれるカテゴリでは非常に良い性能を示すが、全体的には十分な検知精度にまで至っていない。

一方、OCBE モジュールに CBAM を組み込んだ場合、モデルは多くのカテゴリで最良の検知精度を示し、平均値でも最大の値を得た。これは、MFF ブロックにて教師が抽出したマルチスケール特徴を集約した直後に OCAE ブロックが位置する構造により、ノイズ特徴が効果的に緩和された特徴表現を適切に強調可能であるためと考えられる。

表 4.3: CBAM の導入位置による検知精度の定量的評価

(上段 : AUROC (%) ↑, 下段 : AUPRO (%) ↑).

Category	Bread	Cracker	Jelly	Pudding	Soda	Yogurt	Yokan	Avg
Encoder	52.9	83.4	78.2	25.2	57.8	81.7	34.5	59.1
	19.6	56.2	40.7	00.1	20.0	54.1	12.0	27.4
Encoder & Decoder	62.5	<b>90.4</b>	82.4	66.5	42.2	73.2	<b>86.9</b>	72.0
	29.0	<b>72.8</b>	54.2	31.0	04.8	29.1	<b>68.0</b>	41.3
OCBE (Ours)	<b>86.2</b>	84.8	<b>85.3</b>	<b>85.4</b>	<b>81.2</b>	<b>84.8</b>	84.5	<b>84.6</b>
	<b>64.4</b>	55.4	<b>59.9</b>	<b>59.5</b>	<b>44.7</b>	<b>59.4</b>	59.0	<b>57.5</b>

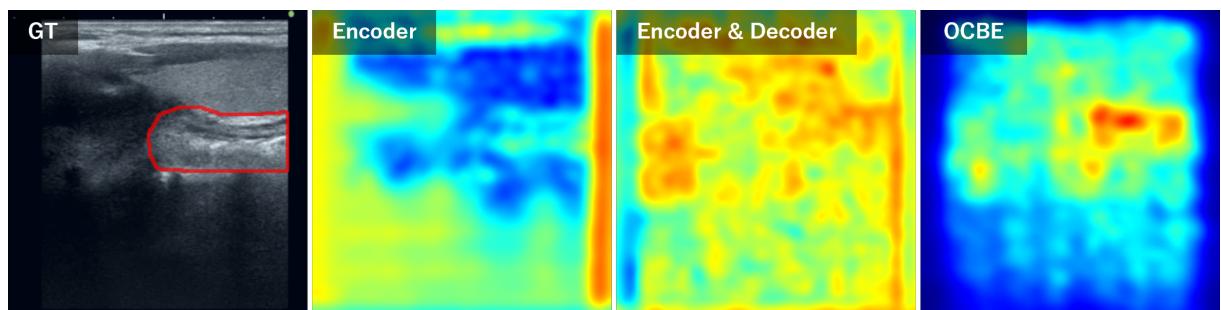


図 4.6: CBAM の導入位置による定性的評価。左から GT、エンコーダ、エンコーダおよびデコーダ、OCBE モジュールへ導入した場合の異常度マップを示す。

### 4.3.3 CBAM の構成に関する検討

本節では、提案手法における適切なモデル構成を実現すべく、CBAM の構成について検討する。提案手法で導入する CBAM の基本構成は、Woo らに従いチャネルアテンション、空間アテンションの逐次的配置を採用している。そこで、各アテンションの適用順序を反転させた場合の検知精度を比較する。表 4.4 に定量的評価、図 4.7 に定性的評価を示す。

Jelly, Soda, Yogurt の 3 種類では、2 つの評価指標において、各アテンションの反転構成が基本構成に対して僅かに検知精度を上回ることが確認された。また、反転構成は基本構成と比較して、表 4.1 で示した競合手法 [6] の検知精度を上回るカテゴリが多い。基本構成では、まずチャネル方向の重要特徴が選択され、その後強調された特徴マップに対して空間方向の重要領域が選択される。一方、反転構成では、空間方向の重要領域を先に選択した後にチャネル方向の重み付けを行う。すなわち、「どこが重要か」を選択したうえで、「その特徴のうちどれが重要か」を調整することになる。ゆえに、反転構成による検知精度の改善は、CBAM が空間的な選択を先に適用することで、食塊領域が小さいまたは輝度差が低い画像において基本構成の CBAM を導入した際に生じたチャネルアテンションによる悪影響が一部緩和されるためと解釈できる。平均値的な観点を考慮すると偏に反転設計が適切ではないが、言い換えれば、カテゴリに応じて各アテンションの適用順序を切り替える動的な設計の可能性を示す知見である。

表 4.4: CBAM の構成による検知精度の定量的評価

(上段: AUROC (%) ↑, 下段: AUPRO (%) ↑).

Category	Bread	Cracker	Jelly	Pudding	Soda	Yogurt	Yokan	Avg
Spatial → Channel	86.1	84.3	<b>85.6</b>	84.7	<b>81.9</b>	<b>85.1</b>	<b>84.6</b>	<b>84.6</b>
	<b>64.7</b>	54.6	<b>60.2</b>	57.8	<b>46.4</b>	<b>60.0</b>	58.9	<b>57.5</b>
Channel → Spatial (Ours)	<b>86.2</b>	<b>84.8</b>	85.3	<b>85.4</b>	81.2	84.8	84.5	<b>84.6</b>
	64.4	<b>55.4</b>	59.9	<b>59.5</b>	44.7	59.4	<b>59.0</b>	<b>57.5</b>

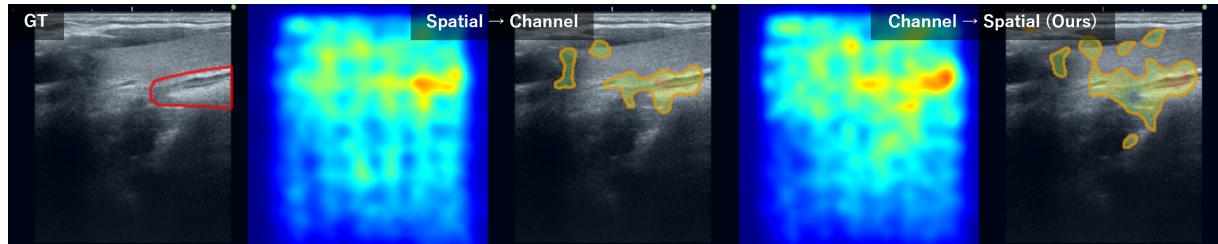


図 4.7: CBAM の構成による定性的評価。左から GT, CBAM の反転構成、基本構成の異常度マップ、予測結果を示す。

#### 4.3.4 ハイパーパラメータおよびバックボーンネットワークに関する検討

本節では、提案手法における適切なモデル構成を実現すべく、チャネルアテンションモジュールにおける MLP の次元縮小率、空間アテンションモジュールにおける畠み込み層のカーネルサイズ、ならびに提案手法のバックボーンネットワークについて検討する。

##### ■ 次元縮小率

表 4.5 に定量的評価を示す。チャネル間関係を利用する手法では、容量を増加させたとしても性能が単調に向上するわけではなく、むしろチャネル間の依存関係に過度に適合してしまうことがある [21]。検知精度と計算量のトレードオフより、提案手法においても多くの CNN モデルと同様に  $r = 16$  が適切であるといえる。

表 4.5: Channel Attention Module における MLP の次元縮小率  $r$  による検知精度の定量的評価（上段：AUROC (%) ↑、下段：AUPRO (%) ↑）。

Category	Bread	Cracker	Jelly	Pudding	Soda	Yogurt	Yokan	Avg
$r = 8$	86.1	84.7	<b>85.3</b>	<b>85.4</b>	<b>81.2</b>	<b>84.8</b>	<b>84.5</b>	<b>84.6</b>
	<b>64.4</b>	<b>55.4</b>	<b>59.9</b>	<b>59.5</b>	<b>44.7</b>	59.4	<b>59.0</b>	<b>57.5</b>
$r = 16$ (Ours)	<b>86.2</b>	<b>84.8</b>	<b>85.3</b>	<b>85.4</b>	<b>81.2</b>	<b>84.8</b>	<b>84.5</b>	<b>84.6</b>
	<b>64.4</b>	<b>55.4</b>	<b>59.9</b>	<b>59.5</b>	<b>44.7</b>	59.4	<b>59.0</b>	<b>57.5</b>
$r = 32$	86.0	84.5	85.2	85.3	81.0	<b>84.8</b>	<b>84.5</b>	84.5
	64.2	55.2	<b>59.9</b>	<b>59.5</b>	<b>44.7</b>	<b>59.5</b>	<b>59.0</b>	57.4

##### ■ カーネルサイズ

表 4.6 に定量的評価を示す。カテゴリ別の検知精度は  $k < 7$  が優勢であるが、平均値の観点では  $k = 7$  が最良となり、モデルの安定性が確認できる。ノイズや異常特徴の誤強調が多いカテゴリでは、受容野が狭い  $k < 7$  の畠み込み層が検知精度の向上に寄与していると考えられる。一方、 $k = 7$  という設定は多くの場合で比較的広範囲の空間的構造を捉えることが可能であるため、提案手法において教師なし食塊検知モデルに CBAM を組み込む目的に適しているといえる。

表 4.6: Spatial Attention Module における畠み込み層のカーネルサイズ  $k$  による検知精度の定量的評価（上段：AUROC (%) ↑、下段：AUPRO (%) ↑）。

Category	Bread	Cracker	Jelly	Pudding	Soda	Yogurt	Yokan	Avg
$k = 3$	84.8	82.7	83.7	<b>85.9</b>	80.3	<b>85.5</b>	83.9	83.8
	62.6	54.4	57.9	<b>59.7</b>	<b>44.8</b>	<b>61.6</b>	58.5	57.1
$k = 5$	84.7	<b>84.9</b>	83.8	84.4	<b>81.4</b>	85.2	<b>84.7</b>	84.2
	62.0	<b>55.8</b>	58.1	57.9	44.2	60.5	<b>59.4</b>	56.8
$k = 7$ (Ours)	<b>86.2</b>	84.8	<b>85.3</b>	85.4	81.2	84.8	84.5	<b>84.6</b>
	<b>64.4</b>	55.4	<b>59.9</b>	59.5	44.7	59.4	59.0	<b>57.5</b>

## ■ バックボーンネットワーク

表 4.7 に定量的評価、図 4.8 に定性的評価を示す。WRN-101-2 は、ボトルネック型残差ブロックを用いた全 101 層の学習可能な重み付き層から構成される。各ステージに配置される残差ブロック数はそれぞれ 3, 4, 23, 3 個であり、合計 33 個の残差ブロックが直列に配置される。

理論的に妥当な結果として、層数が最大である WRN-101-2 は全カテゴリにおいて一貫して高い検知精度を示した。しかし、ResNet-50 と WRN-50-2 との比較では、精度向上が限定的であるカテゴリや、誤差の範囲内ではあるが検知精度が低下するカテゴリが確認された。この結果は、表現力の向上が必ずしも検知性能の改善に直結しない可能性を示唆している。すなわち、WRN-50-2 は ResNet-50 よりも高い表現力を持つものの、その表現力が十分でない場合にはノイズを特徴として抽出してしまう可能性がある。このような特徴表現に対して CBAM を適用すると、抽出されたノイズの誤強調が検知精度の低下を招く要因となり得る。一方、表現力が十分に高い WRN-101-2 ではノイズを特徴として抽出すること自体が抑制され、CBAM が有効に機能し検知精度の向上に貢献したと解釈できる。

以上より、CBAM の効果は T-S モデルを構成するバックボーンネットワークの表現力に強く依存しており、超音波画像に対して表現力が不十分なモデルを使用した場合には、かえって誤強調を引き起こす可能性があることが示唆される。また、計算量について WRN-50 を基準とした場合、ResNet-50 は約 0.85 倍、WRN-101-2 は 1.42 倍であった。結果として、WRN-101-2 は WRN-50 と比較してモデルの深層化に伴い計算コストが僅かに増加する一方、コストに見合った検知精度の改善を得た。

表 4.7: バックボーンネットワークによる検知精度の定量的評価

(上段: AUROC (%) ↑, 下段: AUPRO (%) ↑).

Category	Bread	Cracker	Jelly	Pudding	Soda	Yogurt	Yokan	Avg
ResNet-50	85.1	84.6	84.4	84.4	80.0	85.5	84.6	84.1
	61.9	55.2	61.6	58.6	44.1	61.5	60.6	57.6
WRN-50-2 (Ours)	86.2	84.8	85.3	85.4	81.2	84.8	84.5	84.6
	64.4	55.4	59.9	59.5	44.7	59.4	59.0	57.5
WRN-101-2	<b>86.6</b>	<b>87.2</b>	<b>86.6</b>	<b>85.5</b>	<b>82.3</b>	<b>87.0</b>	<b>86.9</b>	<b>86.0</b>
	<b>64.9</b>	<b>65.0</b>	<b>63.2</b>	<b>59.9</b>	<b>46.8</b>	<b>63.5</b>	<b>62.6</b>	<b>60.8</b>

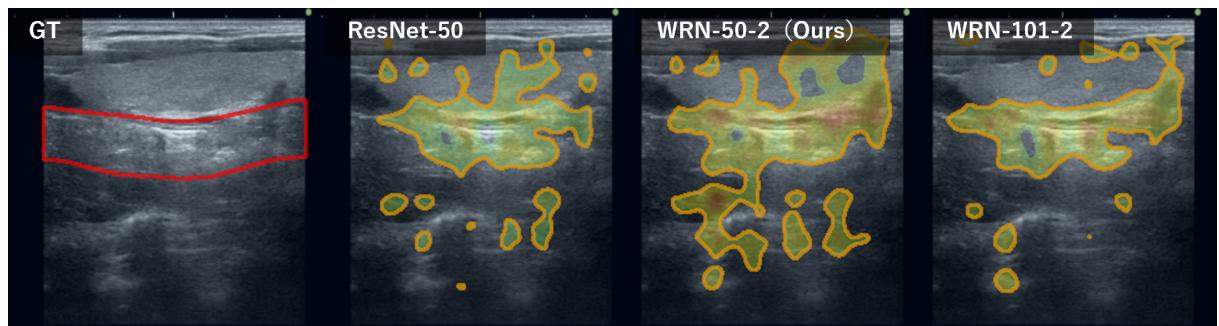


図 4.8: バックボーンネットワークによる定性的評価。左から GT, ResNet-50, WRN-50-2, WRN-101-2 を使用した場合の予測結果を示す。

# 第5章 結論

本研究では、空間的な選択性を付与する注意機構 CBAM を導入した OCAE ブロックを用いた知識逆蒸留に基づく教師なし食塊検知モデルを提案した。

評価実験では、嚥下超音波画像データセットによる画素単位の検知精度および領域一致度の観点から、競合手法に対する提案手法の優位性を示した。提案手法の空間的特徴に着目するモデル構成は、多くのカテゴリで検知精度の改善に貢献することを確認した。

また、アブレーション研究ならびに適切なモデル構成に関する検討では、OCAE ブロックによるチャネル・空間双方向の注意が正常特徴への選択性の安定化に寄与することを示すと共に、CBAM 内部の各アテンションモジュールの配置およびハイパーパラメータについて考察した。

バックボーンネットワークに関する検討では、表現力が十分に高いモデルでは CBAM が有効に機能する一方、不十分なモデルではノイズの誤強調により CBAM が悪影響を及ぼす可能性が示唆された。以上の分析から、CBAM の効果は導入位置と T-S モデルを構成するバックボーンネットワークの表現力に強く依存することが明らかとなった。

今後の展望として、いくつかの点が挙げられる。まず、ハイパーパラメータに関する検討で扱った Spatial Attention Module における畳み込み層のカーネルサイズについて、本研究では平均値的な観点から固定値 ( $k = 7$ ) を採用したが、結果から食塊のカテゴリによって適切なカーネルサイズは異なると推測される。ゆえに、カテゴリ毎に適切なカーネルサイズを採用する動的な設計が実現できれば、検知精度向上に繋がると考えられる。

これに関連して、一部の食塊における検知精度には依然として改善の余地が残されている。特に、Soda は提案手法による精度の向上率こそ大きいものの、検知精度の絶対値は低い水準に留まっている。この要因として、液体は超音波画像における輝度が低いために、ノイズとの輝度差が小さいことが挙げられる。異常度マップから予測結果へ変換する際の閾値には、カテゴリ毎に F1 スコアが最大となる値を採用しているが、単純な閾値設定は誤検知を招く可能性がある。実際に、提案手法の異常度マップにおいて、食塊領域に高スコアが適切に分布している一方、ノイズにも中程度の異常スコアが分布するために局所性が失われているケースが確認された。このような場合、閾値が高スコアと中程度のスコアの間であれば正確な予測結果が得られるため、閾値の最適化は領域一致度の向上に寄与すると考えられる。

また、食塊のカテゴリによらず共通する課題として、嚥下超音波画像データセットにおいて食塊が右側（胃側）に位置するサンプルが多いために、左側（口側）に位置するサンプルでは特に右側のノイズに過敏になり検知精度が低下する傾向が確認された。提案手法による空間的選択性の付与のみでは、データセットによる位置の偏性に起因する問題を十分に抑制することは困難であるため、根本的なノイズの除去に特化したフレームワークの導入や、食塊の物性に起因する輝度差（固体・半固体・液体）を明示的に考慮した特徴抽出が効果的であると考えられる。

さらに、顔表情認識分野においては、CBAM を拡張し、チャネルアテンションと空間

アテンションを並列に適用するハイブリッド型アテンション機構の有用性が報告されている。Liao ら [27] が提案する PH-CBAM は、チャネル分割に基づくスプリットチャネルアテンションを導入し、パラメータ数の増加およびアテンションの逐次的配置に伴う相互干渉を抑制することで性能向上を達成している。これらの知見と、CBAM の構成に関する検討で扱った各アテンションの適用順序により検知精度に差が生じることを踏まえると、本研究に適した形で拡張した PH-CBAM の導入は、表現力の強化という観点から一定の効果が期待される。

今後は、これらの課題に対処しつつ、本手法を摂食嚥下リハビリテーションの臨床応用へ展開することを念頭に、実環境下におけるさらなる性能評価およびモデルの高精度化を目指す。

# 謝辞

本研究を進めるにあたり、指導教員として終始懇切なご指導を頂きました岩手大学 萩原義裕教授、堀田克哉助教に謹んで深謝します。

最後に、本研究について様々な面でご指摘とご協力を頂いた萩原研究室の皆様に感謝致します。

# 参考文献

- [1] 進武幹, 前山忠嗣, and 梅崎俊郎, “嚥下障害,” 口腔・咽頭科, vol.1, pp.93–101, 1989.
- [2] T.N. Doan, W.C. Ho, L.H. Wang, F.C. Chang, N.T. Nhu, and L.W. Chou, “Prevalence and methods for assessment of oropharyngeal dysphagia in older adults: a systematic review and meta-analysis,” Journal of Clinical Medicine, vol.11, no.9, p.2605, 2022.
- [3] I. Fujishima, M. Fujiu-Kurachi, H. Arai, M. Hyodo, H. Kagaya, K. Maeda, T. Mori, S. Nishioka, F. Oshima, S. Ogawa, *et al.*, “Sarcopenia and dysphagia: position paper by four professional organizations,” Geriatrics & Gerontology International, vol.19, no.2, pp.91–97, 2019.
- [4] 辻川敬裕, 榎代茂之, and 平野滋, “嚥下障害と栄養管理・免疫栄養学,” 京都府立医科大学雑誌/京都府立医科大学雑誌編集委員会 編, vol.134, no.6, pp.349–356, 2025.
- [5] 藤島一郎, “嚥下障害と誤嚥・咽頭残留の病態及びその対処法,” 日本バイオレオロジー学会誌, vol.20, no.2, pp.2–9, 2006.
- [6] Q. Gao, Y. Hagihara, M. Sasaki, C. Gu, and K. Hotta, “Unsupervised food bolus detection for ultrasound images via reverse distillation,” Proceedings of the 2025 IEEE 14th Global Conference on Consumer Electronics (GCCE), IEEE, IEEE, sep 2025.
- [7] S. Woo, J. Park, J.Y. Lee, and I.S. Kweon, “Cbam: Convolutional block attention module,” Proceedings of the European conference on computer vision (ECCV), pp.3–19, 2018.
- [8] Q. Gao, Y. Hagihara, M. Sasaki, and K. Hotta, “Attention-guided food bolus segmentation in ultrasound imaging for dysphagia rehabilitation,” IEEJ Transactions on Electrical and Electronic Engineering, vol.20, no.11, pp.1757–1765, 2025.
- [9] C. Han, L. Rundo, K. Murao, T. Noguchi, Y. Shimahara, Z.Á. Milacski, S. Koshino, E. Sala, H. Nakayama, and S. Satoh, “Madgan: Unsupervised medical anomaly detection gan using multiple adjacent brain mri slice reconstruction,” BMC bioinformatics, vol.22, no. Suppl 2, p.31, 2021.
- [10] D. Bhattacharya, F. Behrendt, B.T. Becker, D. Beyersdorff, E. Petersen, M. Petersen, B. Cheng, D. Eggert, C. Betz, A.S. Hoffmann, *et al.*, “Unsupervised anomaly detection of paranasal anomalies in the maxillary sinus,” Medical Imaging 2023: Computer-Aided Diagnosis, pp.291–296, SPIE, 2023.

- [11] B. Schölkopf, R.C. Williamson, A. Smola, J. Shawe-Taylor, and J. Platt, “Support vector method for novelty detection,” Advances in neural information processing systems, vol.12, 1999.
- [12] G. Hinton, O. Vinyals, and J. Dean, “Distilling the knowledge in a neural network,” arXiv preprint arXiv:1503.02531, 2015.
- [13] L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S.A. Siddiqui, A. Binder, E. Müller, and M. Kloft, “Deep one-class classification,” International conference on machine learning, pp.4393–4402, PMLR, 2018.
- [14] D. Gong, L. Liu, V. Le, B. Saha, M.R. Mansour, S. Venkatesh, and A.v.d. Hengel, “Memorizing normality to detect anomaly: Memory-augmented deep autoencoder for unsupervised anomaly detection,” Proceedings of the IEEE/CVF international conference on computer vision, pp.1705–1714, 2019.
- [15] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, “Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings,” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp.4183–4192, 2020.
- [16] H. Deng and X. Li, “Anomaly detection via reverse distillation from one-class embedding,” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp.9737–9746, 2022.
- [17] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” Proceedings of the IEEE conference on computer vision and pattern recognition, pp.770–778, 2016.
- [18] F. Tung and G. Mori, “Similarity-preserving knowledge distillation,” Proceedings of the IEEE/CVF international conference on computer vision, pp.1365–1374, 2019.
- [19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” Advances in neural information processing systems, vol.30, 2017.
- [20] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, “Residual attention network for image classification,” Proceedings of the IEEE conference on computer vision and pattern recognition, pp.3156–3164, 2017.
- [21] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” Proceedings of the IEEE conference on computer vision and pattern recognition, pp.7132–7141, 2018.
- [22] M.D. Zeiler and R. Fergus, “Visualizing and understanding convolutional networks,” European conference on computer vision, pp.818–833, Springer, 2014.
- [23] S. Zagoruyko and N. Komodakis, “Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer,” arXiv preprint arXiv:1612.03928, 2016.

- [24] S. Zagoruyko and N. Komodakis, “Wide residual networks,” arXiv preprint arXiv:1605.07146, 2016.
- [25] F.S. Nahm, “Receiver operating characteristic curve: overview and practical use for clinicians,” Korean journal of anesthesiology, vol.75, no.1, pp.25–36, 2022.
- [26] Y. Lin and X. Li, “Back to the metrics: Exploration of distance metrics in anomaly detection,” Applied Sciences, vol.14, no.16, 2024.
- [27] L. Liao, S. Wu, C. Song, and J. Fu, “Ph-cbam: A parallel hybrid cbam network with multi-feature extraction for facial expression recognition,” Electronics, vol.13, no.16, p.3149, 2024.