注意機構を用いた知識逆蒸留による嚥下機能評価の検討

鈴木 晴仁[1,a)]　萩原 義裕[1,b)]　堀田 克哉[1,c)]

概要：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．教師なし異常検知を用いた手法は，無数に存在し得る嚥下パターンを網羅的にラベル付けする必要がなく，正常サンプルのみで学習できる反面，検知精度に改善の余地がある．本研究では，既存の教師なし食塊検知モデルに注意機構を組み込むことを提案する．嚥下超音波画像データセットによる広範な実験を通して，空間的な選択性を付与する提案手法の有効性を確認した．

キーワード：CBAM，知識逆蒸留，教師なし異常検知

1. はじめに

嚥下障害とは，飲食物を口腔から胃まで運搬する一連の運動における障害をいう．嚥下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．嚥下障害発症の主な原因は，脳血管疾患や神経変性疾患である．しかし，近年では高齢化に伴い，全身性のサルコペニアや低栄養を併存した嚥下障害患者が増加している [1]．これらの影響は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 [2]．

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，咽頭残留を検出することが重要である [3]．嚥下機能評価のゴールドスタンダードは嚥下造影検査（VideoFluoroscopic examination of swallowing: VF）であるが，設備や侵襲性，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．また，医師による診断は VF 画像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．

本研究では，空間的な選択性を付与する注意機構 [4] を導入した知識逆蒸留に基づく食塊検知モデルを提案する．本手法は，VF に代わる非侵襲的かつ定量的な嚥下機能の評価手法である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．実験では，嚥下超音波画像データセットにおける画素単位の検知精度を評価することで，提案手法の有効性を示す．

[1] 岩手大学
4-3-5, Ueda, Morioka, Iwate 020–8551, Japan
[a)] s0622032@iwate-u.ac.jp
[b)] dhag@iwate-u.ac.jp
[c)] hotta@iwate-u.ac.jp

2. 関連研究

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，特に咽頭における食塊の検出に焦点を当てる．

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 [5] 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いた食塊検知モデルを提案した．本手法は，予測に無関係な領域における特徴活性化を抑制するアテンションゲートにより，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し，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．

食塊形状の多様性に対処するため，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．本手法では，嚥下動作を伴わず食塊を含まない食道状態を正常と定義し，その特徴分布を学習する．推論時には，学習した正常特徴表現の分布から逸脱する食塊を含む特徴を異常として検知する．このような教師なし学習に基づく手法は，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 [7], [8].

3. 提案手法

本研究は，先行研究 [6] に着想を得て，OCC に基づく食塊検知手法を提案する．はじめに，提案手法が採用する知識逆蒸留の予備知識について述べる．次に，提案手法である注意機構を用いた知識逆蒸留について説明する．

3.1 予備知識：知識逆蒸留

知識蒸留 [9] は，生徒は教師と同様または類似した NN を採用し，入力データから教師の特徴活性化を模倣するフレームワークである．教師なし異常検知の一クラス蒸留 [10] においては，生徒が異常サンプルに対して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．しかし，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．この原因は，教師と生徒モデル（T-S モデル）の類似性および知識蒸留時にデータフローが同一であることに起因する．上記の問題に対処するため，Deng ら [11] 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（図 1）．知識逆蒸留において，教師は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とし，学習中に教師の挙動を模倣することを目指す．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できる．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．推論時には，マルチスケール特徴に基づく蒸留を用いることで，局所的かつ領域的な異常を示す．

One-Class Bottleneck Embedding: 知識逆蒸留 [11] では，ボトルネック埋め込みのコンパクト性が異常検知の性能に重要であるため，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 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．MFF ブロックの畳み込み層および OCE ブロックの ResBlock は学習可能であり，生徒と共同で正常サンプルにより最適化される．異常特徴を正常パターン上の微細な摂動とみなした場合，このコンパクトな埋め込みは情報ボトルネックとして機能し，異常特徴の生徒への伝播を防ぐ．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．

知識蒸留損失: 知識逆蒸留では，知識蒸留損失により OCBE モジュールと生徒デコーダが共同で学習される．具体的には，入力データ I から一クラスボトルネック埋め込み空間への射影を ϕ とすると，教師エンコーダと生徒デコーダの

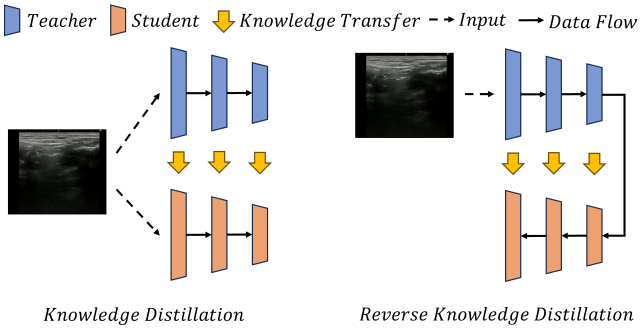

図 1 知識蒸留の概要．左: 知識蒸留，右: 知識逆蒸留．

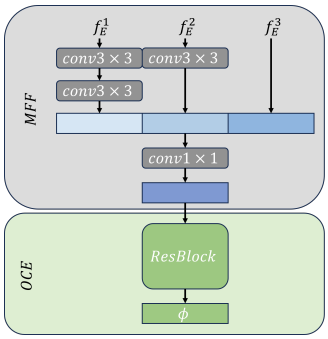

図 2 OCBE モジュールの概要．

対応する活性化は次のようになる:

$$f_E^k = E^k(I), \tag{1}$$

$$f_D^k = D^k(\phi). \tag{2}$$

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 はチャネル数，高さ，幅を示す．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:

$$M^k(h,w) = 1 - \frac{(f_E^k(h,w))^T \cdot f_D^k(h,w)}{\|f_E^k(h,w)\| \|f_D^k(h,w)\|}. \tag{3}$$

マルチスケール蒸留を考慮し，式 (3) のスカラー損失関数は全スケールの異常マップを加算することで定義される:

$$\mathcal{L}_{KD} = \sum_{k=1}^{K} \left\{ \frac{1}{H_k W_k} \sum_{h=1}^{H_k} \sum_{w=1}^{W_k} M^k(h,w) \right\}. \tag{4}$$

ここで，K は考慮する特徴階層数を表す．異常スコアの算出時には，前述の式 (3) に従い，各層ごとにピクセル単位の再構成誤差を計算し，K 枚の異常マップを得る．最終的に，画像全体の異常マップ S_{AL} は，M^k をバイリニア補間 Ψ により画像サイズにアップサンプリングしたのち，ピクセル単位で全スケールを加算する:

$$S_{AL} = \sum_{i=1}^{L} \Psi(M^i). \tag{5}$$

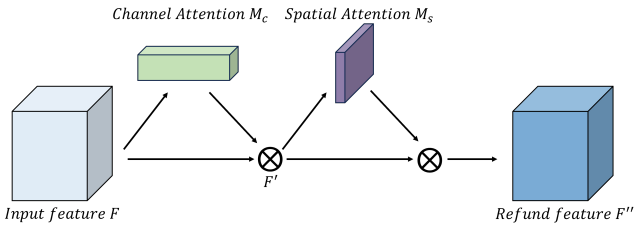

図 3 CBAM の概要.

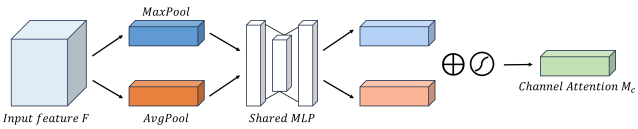

図 4 Channel Attention Module の概要.

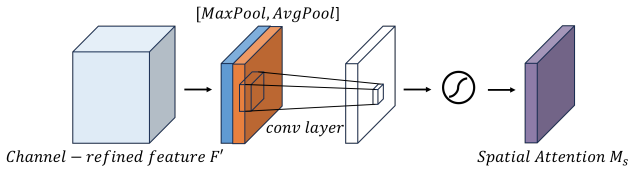

図 5 Spatial Attention Module の概要.

3.2 注意機構を用いた知識逆蒸留

本研究では，T-S モデルにおける OCBE モジュールが情報を圧縮する際に，注目すべき特徴を強調するためにチャネルアテンションと空間アテンションを併用する注意機構である Convolutional Block Attention Module (CBAM) [4] を導入する．

CBAM: CBAM は，図 3 に示すように，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．中間特徴マップ $F \in \mathbb{R}^{C\times H\times W}$ に対して，1 次元のチャネルアテンションマップ $M_c \in \mathbb{R}^{C\times 1\times 1}$ と 2 次元の空間アテンションマップ $M_s \in \mathbb{R}^{1\times H\times W}$ を順次推定する．すなわち，最終的なアテンションの出力 F'' は次式で計算される:

$$F' = M_c(F) \otimes F, \tag{6}$$

$$F'' = M_s(F') \otimes F'. \tag{7}$$

ここで $\otimes$ は要素ごとの積を示し，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 [4] 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．

Channel Attention Module: 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（図 4）．特徴マップの各チャネルは一種の特徴検出器 [12] とみなせるため，チャネルアテンション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．具体的には，特徴マップに対して平均プーリングと最大プーリングを適用し，F^c_{avg}，F^c_{max} 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．これらは 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 を通して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 $M_c \in \mathbb{R}^{C\times 1\times 1}$ 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:

$$\begin{aligned} M_c(F) &= \sigma\left(\mathrm{MLP}(\mathrm{AvgPool}(F)) + \mathrm{MLP}(\mathrm{MaxPool}(F))\right) \\ &= \sigma\left(W_1(W_0(F^c_{\mathrm{avg}})) + W_1(W_0(F^c_{\mathrm{max}}))\right) \end{aligned} \tag{8}$$

ここで σ は活性化関数，r は MLP における次元縮小率を示す．また，$W_0 \in \mathbb{R}^{C/r\times C}, W_1 \in \mathbb{R}^{C\times C/r}$ は，それぞれ MLP における特徴変換行列を示す．

Spatial Attention Module: 空間アテンションマップは，図 5 に示すように，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．具体的には，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 F^s_{avg}，$F^s_{max} \in \mathbb{R}^{1\times H\times W}$ を得る．これらの特徴マップの連結後，畳み込み層を通して 2 次元の空間アテンションマップ $M_s(F) \in \mathbb{R}^{H\times W}$ を生成する:

$$\begin{aligned} M_s(F) &= \sigma\left(f^{7\times 7}\left([\,\mathrm{AvgPool}(F); \mathrm{MaxPool}(F)\,]\right)\right) \\ &= \sigma\left(f^{7\times 7}\left([\,F^s_{\mathrm{avg}}; F^s_{\mathrm{max}}\,]\right)\right) \end{aligned} \tag{9}$$

ここで σ は活性化関数，$f^{7\times 7}$ はフィルタサイズ 7×7 の畳み込み演算を表す．

CBAM を用いた知識逆蒸留に基づく T-S モデル: 提案手法は，教師エンコーダ E と CBAM を導入した学習可能な OCBE モジュール ϕ，生徒デコーダ D で構成される（図 6）．また，教師モデル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，特徴 x を低次元空間に埋め込み，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

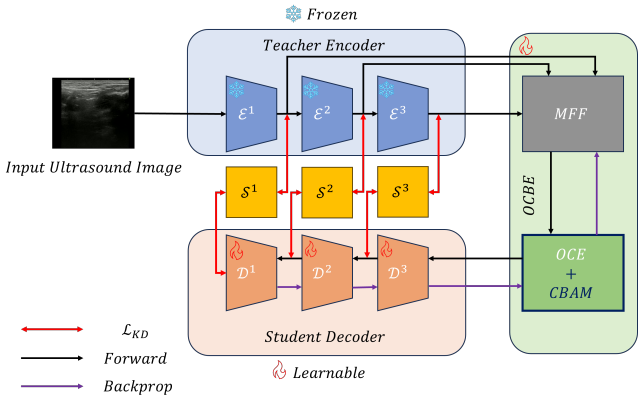

図 6 提案手法の概要.

表 1　嚥下超音波画像データセット [5] における食塊検知精度の評価（AUROC ↑ / PRO ↑）.

Category	Bread	Cracker	Jelly	Pudding	Soda	Yogurt	Yokan	Avg
AD [6]	84.7 / 63.2	**82.8** / **54.9**	83.7 / 58.3	84.1 / 57.0	79.4 / 42.7	**85.2** / **60.7**	**84.8** / **60.0**	83.4 / 56.7
Ours	**85.2** / **63.4**	**82.8** / 54.4	**84.3** / **58.9**	**84.4** / **58.5**	**80.0** / **43.7**	84.8 / 60.4	83.8 / 58.2	**83.6** / **57.1**

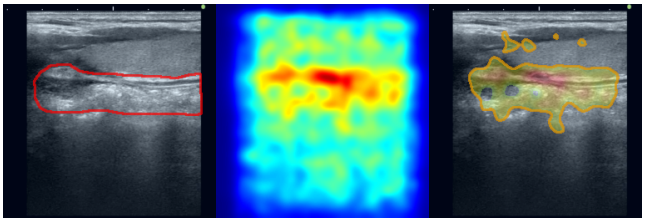

図 7　提案手法の定性的結果．左から GT，異常度マップ，予測結果を示す.

デコーダ D は，z を入力として教師の特徴を再構成するように学習を行う．これにより，推論時に異常サンプルが入力された場合，生徒デコーダによる再構成誤差を異常として検知することが可能である.

4.　実験

本節では，嚥下超音波画像データセット [5] において提案手法を評価する．嚥下超音波画像データセットは，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．競合手法と同一条件下で定量的比較を実施するために，Gao ら [6] に従い，652 枚を学習，1743 枚をテストとして用いる．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/PRO を用いる.

4.1　実験結果

表 1 に定量的評価，図 7 に定性的評価を示す．各評価指数の列において，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 [6] を上回ることが確認できる．提案手法は，CBAM を組み込んだ OCBE 内でチャネルおよび空間方向の注意を与えることで，図 7 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，OCBE のみで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できる．また，提案手法は，テスト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．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，結果的に検知精度の向上に寄与していると考えられる.

4.2　分析

4.2.1　アブレーション研究

提案手法における各モジュールの有効性を検証するために，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表 2 に示す．また，各アブレーションにおける異常度マップを図 8 に示す．提案手法から CBAM（Channel + Spatial Attention）を除いた場合，検知精度の低下が確認できる．特に Spatial Attention を除いた場合の PRO 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．また，すべてのモジュールを除いた場合，2 つの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モデルは潜在特徴の冗長性が高いために，生徒モデルが正常パターンの本質的な特徴を復元することを妨げることが要因と考えられる．実際に図 8 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復元が困難となることが確認できる．そのため，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ことが確認できる.

表 2　アブレーション研究.

OCBE	Channel	Spatial	AUROC (%) ↑	PRO (%) ↑
			80.1	46.0
✓			83.4	56.7
✓	✓		**83.6**	56.8
✓	✓	✓	**83.6**	**57.1**

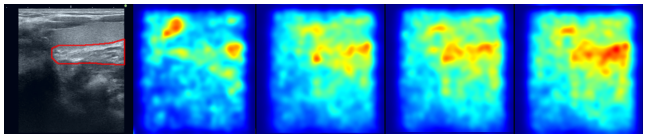

図 8　アブレーション研究における異常度マップ．左から GT，すべてのモジュールの除去，CBAM の除去，Channel Attention のみ除去，提案手法の結果を示す.

4.2.2　CBAM の挿入位置

提案手法はエンコーダ，OCBE モジュール，デコーダが全て ResNet 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 を挿入した定量的評価を表 3 に示す．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性能を示した．これは，エンコーダ側で CBAM がノイズを誤強調することで，OCBE モジュールおよび生徒デコーダに伝播することが原因と考えられる．エンコーダとデコーダの両モデルに CBAM を組み込んだ場合は，前述の不適切な誤差が緩和されるために食塊検知精度の向上が確認できる．しかし，OCBE モジュールに CBAM を組み込んだ場合と比べて，大きく検知精度が低下している．これは，MFF 後に CBAM 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できるためと考えられる.

表 3 CBAM の挿入位置による検知精度の評価.

Encoder	Decoder	OCBE	AUROC (%) ↑	PRO (%) ↑
✓			59.1	27.4
✓	✓		72.0	41.3
		✓	**83.6**	**57.1**

5. まとめ

本研究では，注意機構 CBAM を導入し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．評価実験の結果，提案手法の空間的な選択性を付与するモデル構成は，多くのカテゴリで検知精度の改善に貢献することが確認された．今後は，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．

参考文献

[1] Fujishima, I., Fujiu-Kurachi, M., Arai, H., Hyodo, M., Kagaya, H., Maeda, K., Mori, T., Nishioka, S., Oshima, F., Ogawa, S. et al.: Sarcopenia and dysphagia: position paper by four professional organizations, *Geriatrics & Gerontology International*, Vol. 19, No. 2, pp. 91–97 (2019).

[2] Tsujikawa, T., Mukudai, S. and Hirano, S.: Dysphagia and Nutritional Management: An Immunonutritional Perspevctive, *JOURNAL OF KYOTO PREFECTUAL UNIVERSITY OF MEDICINE*, Vol. 134, No. 6, pp. 349–356 (2025).

[3] Fujishima, I.: Pathophysiology and management of dysphagia, aspiration and pharyngeal residue, *NPO Japanese Society of Biorheology*, Vol. 20, No. 2, pp. 52–59 (2006).

[4] Sanghyun, W., Jongchan, P., Joon-Young, L. and In So, K.: CBAM: Convolutional Block Attention Module, *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 3–19 (2018).

[5] Gao, Q., Hagihara, Y., Sasaki, M. and Hotta, K.: 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).

[6] Gao, Q., Hagihara, Y., Sasaki, M., Gu, C. and Hotta, K.: Unsupervised Food Bolus Detection for Ultrasound Images via Reverse Distillation, *Proceedings of the 2025 IEEE 14th Global Conference on Consumer Electronics (GCCE)*, IEEE, IEEE (2025).

[7] Han, C., Rundo, L., Murao, K., Noguchi, T., Shimahara, Y., Milacski, Z. c., Koshino, S., Sala, E., Nakayama, H. and Satoh, S.: MADGAN: unsupervised medical anomaly detection GAN using multiple adjacent brain MRI slice reconstruction, *BMC Bioinformatics*, Vol. 22, No. Suppl 2, p. 31 (2021).

[8] Bhattacharya, D., Behrendt, F., Becker, B. T., Beyersdorff, D., Petersen, E., Petersen, M., Cheng, B., Eggert, D., Betz, C., Hoffmann, A. S. and Schlaefer, A.: Unsupervised Anomaly Detection of Paranasal Anomalies in the Maxillary Sinus, *Proceedings of SPIE Medical Imaging: Image Processing (SPIE 12465)*, Vol. 12465 (2023).

[9] Hinton, G., Vinyals, O. and Dean, J.: Distilling the Knowledge in a Neural Network, *NIPS Deep Learning and Representation Learning Workshop* (2015).

[10] Bergmann, P., Fauser, M., Sattlegger, D. and Steger, C.: Uninformed Students: Student-Teacher Anomaly Detection With Discriminative Latent Embeddings, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4183–4192 (2020).

[11] Hanqiu, D. and Xingyu, L.: Anomaly Detection via Reverse Distillation from One-Class Embedding, *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9727–9736 (2022).

[12] Zeiler, M. D. and Fergus, R.: Visualizing and Understanding Convolutional Networks, *European Conference on Computer Vision (ECCV)*, Springer, pp. 818–833 (2014).