2ZC-08

深層学習を用いた高精細魚種判別モデルの開発に関する一考察

中野 雄太† 長谷川 達人†

福井大学大学院工学研究科†

1. はじめに

水産庁がすすめる資源調査では，持続可能な漁業を確保するために，漁獲物の詳細情報を収集している．この作業は一般に水産試験場の職員が手作業で行っており，作業の効率化が求められている．一方で，世界中には 36,000 種以上，日本近海には 4,400 種以上の魚が存在している．資源調査の対象魚種は 192 種に限定されているが，それでも多くの魚種を識別する必要がある．また，192 種の中にも見た目の似た魚が多く，かつ計測可能なサンプル数が限定的であることから，正確な魚種の詳細分類は決して容易なタスクではない．

これらの問題解決に向けて，正確な魚種の詳細画像分類（FGIR：Fine-Grained Image Recognition）により資源調査を自動化するシステムを我々は開発している．本研究では，深層学習を用いた FGIR モデルと基盤モデルを用いた手法を提案する．魚種分類における FGIR モデルの有効性を評価した上で，基盤モデルによる背景除去が推定精度に与える影響を明らかにする．本手法により，魚種の詳細分類の推定精度を向上させ，正確な資源調査の効率化を図ることを目的とする．

2. 関連研究

FGIR モデルの関連事例として，MMAL-Net[1], CMAL-Net[2], HERBS[3]がある．MMAL-Net は，3 つの Convolutional Neural Network（CNN）で構成されており，それぞれの CNN は画像内の物体位置の予測，情報量の多いパーツ領域の提案，提案されたパーツ領域の特徴抽出という役割を担っている．CMAL-Net は，全体的な性能を向上させるために，CNN の異なる層が互いに学習することで，識別可能な領域への注視を促している．HERBS は，多様な特徴学習の促進，背景ノイズの抑制という 2 つの大きな役割がある．

基盤モデルの一種として，Grounding DINO がある．これは，zero-shot で物体検出が可能であり，入力に画像と対象物体の名称（プロンプト）を受け取り，出力として画像内の対象物体位置をバウンディングボックス（bbox）で返す．

上述の FGIR モデルは，CUB-200-2011（鳥）や Stanford Cars（自動車）のようなデータセットを用いて有効性を検証しているが，魚種分類に有効かは明らかになっていない．また，上述の FGIR モデルは，背景部分の除去や背景というノイズの抑制といった背景部分の情報を取り除く働きを有するが，モデル学習時に教師なしで実施されているため，別モデルがこの役割を担うことで全体の性能を改善できる可能性がある．

3. 提案手法

本研究では，魚種分類において FGIR モデルの適用による有効性を評価する．さらに，Grounding DINO を用いたデータセットの前処理手法を提案し，推定精度に与える影響を明らかにする．

まず，モデル学習前に使用するデータセットに対して Grounding DINO を用いて，魚領域を切り取った画像データセットを新たに作成する．このとき，Grounding DINO に与えるプロンプトは“fish”とし，1 枚の画像に対して魚領域の複数検出時には，最も確信度の高い魚領域を採用する．

次に，モデル学習時では FGIR モデルとして，MMAL-Net，CMAL-Net，HERBS を用いて，魚種分類における FGIR モデルの有効性を評価する．MMAL-Net は物体の位置を教師なしで学習し，その領域を切り取る点で，Grounding DINO を用いたデータセットの前処理手法と似た部分があり，教師なしによる領域の切り取りを行う MMAL-Net と，Grounding DINO を用いた魚領域検出能力の差による影響を明らかにするために採用した．また，最先端な FGIR モデルである CMAL-Net や HERBS は，魚種分類が有効的に働くのかを明らかにするために採用した．

4. 評価実験

評価には，Zhuang らが設計した大規模魚データセット WildFish[5]を用いる．WildFish は 1000 種 54459 枚の画像で構成されているが，解像度が統一されていないため，本研究の評価実験では 448×448[px]にリサイズして用いる．このとき，アスペクト比を固定した上でリサイズし，余白部分に対しては黒埋めを行う．今回，計算時間の都合上，各データセット内の画像枚数を全体の 4 分の 1（13614 枚）に絞って訓練および検証を行う．

本研究では，オリジナルを WildFish（i. base）とした上で，前処理を適用した 4 つのデータセット（ii. dino_35，iii. dino_50，iv. dino_60，v. dino_area）を作成した．dino_*の*は魚領域検出する際の確信度の閾値[%]を表している．dino_35 は正常な検出例が多くある中，対象の魚を検出で

Development of High-Resolution Fish Species Classification Model Using Deep Learning: A Study
† Yuta Nakano, Tatsuhito Hasegawa, Graduate School of Engineering, Unversity of Fukui

きない場合がいくつか確認された．特徴として，背景部分を誤検出する傾向がある．また，正常な検出例が多かったが，背景部分が最も確信度が大きくなる事例もある．共通点として，確信度が最も大きい領域が背景部分であり，かつ領域サイズが小さい．これらの問題に対して，dino_50，dino_60 を作成した．また，dino_35 に対して，魚領域の面積が224×224=50176以上のものに限定する dino_area を作成した．以上の 4 つのデータセットは，各条件により検出された魚領域のうち，最も確信度の高い魚領域の bbox 情報をもとに領域の切り取りを行う．魚領域を検出できなかった画像に対しては base 画像を用いる．

評価実験では，①ResNet50，②MMAL-Net，③CMAL-Net，④SwinTransformer，⑤HERBS を比較対象として検証を行う．各モデルは，ImageNet で事前訓練された重みをファインチューニングした．最適化関数は SGD，初期学習率は 1e-3，エポック数は 100 とし，50 エポックと 80 エポックで，学習率を 0.1 倍にするスケジューリングを行った．また，各モデル訓練時に使用するデータ拡張として，標準化，左右反転，明るさおよびコントラストの変化を行う．全体の評価は，各データセットで訓練したモデルに対して base による推論を行い，評価指標として Accuracy を用いる．

5. 結果・考察

実験の結果を表 1 に示す．まず，魚種分類における FGIR モデルの有効性について，②と③は①と比較，⑤は④と比較すると，それぞれ精度向上がみられるため，魚種分類においても FGIR モデルは有効に働くことがわかる．次に，Grounding DINO による背景除去が推定精度に与える影響について，i と Grounding DINO を適用したデータセット（ii，iii，iv，v）を各モデルで比較する．各モデルに共通して，Grounding DINO を適用した場合に i よりも精度低下がみられた．i の画像から領域を切り取ることによる，画像全体の情報量の減少が精度低下に影響している．別途 Grounding DINO による推論を行ったが，背景除去を行わない方が良いことが分かった．したがって，データセットの特性によっては Grounding DINO による背景除去が推定精度に必ずしも有効的に働かないことが分かった．また，Grounding DINO は魚に特化していないため，魚領域検出性能にも限界があると考えられる．次に，i に対して ii～v をデータ拡張として用いた場合（i+ii～i+v）について，それぞれのデータセット単体で学習を行った場合に比べて，精度向上がみられた．これは各モデル訓練時に，左右反転のような一般的なデータ拡張を使う前提でいたため，単純にデータ量増加による要因が大きいが，Grounding DINO による前処理手法を用いたデータ拡張により，モデルの汎化性能向上に貢献したと考えられる．

6. まとめ

本稿では，深層学習を用いた FGIR モデルと基盤モデルを用いた手法を提案した．この提案では，魚種分類における FGIR モデルの有効性を評価したした上で，基盤モデルによる背景除去が推定精度に与える影響について，検証データを用いて評価した．魚種分類においても FGIR モデル適用の有効性を示すことができたが，Grounding DINO による背景除去は推定精度向上につながらなかった．これはデータセットの特性に依存すると考えられる．しかし，データ拡張として Grounding DINO による前処理手法を用いることで，モデルの汎化性能向上に貢献することを明らかにした．今後は他の大規模魚データセットでも検証を行っていきたい．

謝辞

本研究は，JST ACT-X（JPMJAX20AJ）の支援を受けたものであり，ここに感謝の意を表する．

参考文献

[1] Zhang, F., et al.: Multi-branch and Multi-scale Attention Learning for Fine-Grained Visual Categorization, In Proc. of the MultiMedia Modeling (MMM), (2021).

[2] Liu, D., et al.: Learn from each other to Classify better: Cross-layer mutual attention learning for fine-grained visual classification, Pattern Recognition, (2023).

[3] Chou, P., et al.: Fine-grained Visual Classification with High-temperature Refinement and Background Suppression, arXiv preprint arXiv:2303.064422 (2023).

[4] Liu, S., et al.: Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection, arXiv preprint arXiv:2303.05499 (2023).

[5] Zhuang, P., et al.: WildFish: A Large Benchmark for Fish Recognition in the Wild, In Proc. of the 26th ACM international conference on Multimedia, (2020).

表 1　評価実験の結果[%]

	i	ii	iii	iv	v	i + ii	i + iii	i + iv	i + v
①ResNet50	67.26	54.09	57.59	58.94	63.61	81.37	81.76	81.98	82.54
②MMAL-Net	72.36	63.98	65.79	67.24	68.85	84.65	85.09	85.53	86.31
③CMAL-Net	76.74	63.61	65.60	68.58	71.06	86.17	87.10	86.88	87.44
④SwinTransformer	82.30	79.90	80.19	80.41	80.66	89.91	89.72	90.06	90.33
⑤HERBS	82.57	80.29	81.17	80.36	81.78	90.62	90.52	90.28	90.67