H-040

少量のアノテーション画像と基盤モデルによる藻場の推定
Detection of Seagrass Meadows Using Foundation Models and a Small Number of Annotated Images

大下 真之介† 飯山 将晃†
Shinnosuke Ohshita Masaaki Iiyama

1. はじめに

近年、温室効果ガスの排出に伴う地球温暖化現象が懸念されており、温室効果ガスの排出量と吸収量を均衡させるカーボンニュートラルが現在推進されている。地球上の炭素のうち海洋生態系に取り込まれた炭素をブルーカーボンと呼び、ブルーカーボンがカーボンニュートラルの達成のために注目されており、ブルーカーボンの生態系の把握が重要になっている。このような背景のもと、水中カメラで撮影された画像から藻場の分布を自動的に計測する技術が求められている。本研究では、画像認識技術のひとつであるセマンティックセグメンテーションに着目し、これを用いて水中画像から藻場の推定を行う手法を提案する。

水中画像から藻場を推定する従来の研究には、主に 2 つの課題が存在する。1 つ目は画像劣化である。水中画像はその特性上、光の吸収や散乱の影響を受けやすく、視認性が低下しやすい。具体的には、画像全体が青みを帯び、藻場と海底の色のコントラストが低くなる。このような視覚的な劣化は、モデルによる藻場の正確な推定を妨げる要因となる。2 つ目の課題は、従来の手法が教師あり学習に基づいてモデルを構築しており[1,2]、大量のアノテーション付き画像が必要となる点である。一般的に、アノテーションデータの取得には金銭や時間、専門知識を要する。特に藻場領域推定においては専門家の知識が必要であり十分なアノテーション画像を用意することが難しい。

そこで本研究では、まず画像復元を通じて藻場画像の劣化を改善し、視認性の向上を図る。さらに、基盤モデルである Contrastive Language-Image Pre-training[3]（CLIP）を活用して生成された疑似ラベルを用いた半教師あり学習を導入し、アノテーションコストを抑えたモデルを構築する。

2. 関連研究

本研究に関連する先行研究として、水中画像からの藻場の推定に関する研究について述べる。

2.1 水中画像からの藻場の推定

水中画像からセマンティックセグメンテーションを用いて藻場を推定する手法が提案されている。Reus[1]は水中画像から海藻の被度の自動推定を目的にスーパーピクセルベースの分類とパッチベースの分類の 2 つの手法を提案している。スーパーピクセルベースによる手法では、入力画像から生成されたスーパーピクセルを長方形のパッチに変換し、その長方形のパッチから特徴量抽出器より特徴量を抽出した後、2 クラス分類器によってパッチが海藻か背景かを予測する。そして、予測されたパッチと生成されたスーパーピクセルを照合させる。パッチベース分類による手法では、入力画像から長方形のパッチを生成し、特徴量抽出器より特徴量を抽出した後、分類器でパッチが海藻か背景かを予測する。一方、Weidmann[2]は、深層学習を用いた水中画像からの藻場推定手法を提案しているが、十分な枚数のアノテーション画像が存在していることを前提としている。

また、基盤モデルである CLIP を水中画像から藻場の推定に活用した研究がある[4]。画像をパッチに分割した後、藻場と背景に対応するプロンプトを複数個ずつ用意し、各パッチ画像を CLIP で画像分類する。この研究では、パッチ画像ごとに藻場か背景かを 2 値分類するため、細かい境界線や部分的な藻場の存在を正確に捉えにくいという問題がある。また、パッチのサイズによっては藻場と背景が混在する領域が適切に分類されない可能性があり、セマンティックセグメンテーションと比べて局所的な精度が低下する問題がある。

3. 提案手法

画像復元を導入することで、水中画像特有の画像劣化の解決を図る。また、アノテーション画像取得のコストを抑えるために、基盤モデル CLIP を活用し生成した疑似ラベルを用いる半教師あり学習を用いる。図 1 に提案手法の全体像を示す。

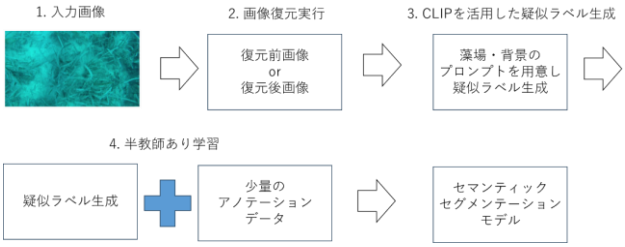

図 1　提案手法の概要図

3.1 画像復元

図 2 に本研究で用いる画像復元の概要図を示す。本研究では、EUVP データセット[5]を用いて、Feature Fusion を構築した（図 2 の FFA①）。その後、藻場推定に用いる水中画像に対して、当該モデルを用いた画像復元処理を実施する。復元処理により、画像によっては正常な画像復元が行われない場合が存在する。こうした低品質な復元画像をセマンティックセグメンテーションモデルの学習に使用する

† 滋賀大学大学院データサイエンス研究科
Graduate School of Data Science, Shiga University

と、藻場推定精度の低下を招くおそれがある。そこで、画像復元に失敗した画像を自動的に検出する手法を導入する。まず、EUVP データセットを用いて、復元画像を元画像へと変換する変換モデルを構築する（図 2 の FFA②）。次に、藻場推定用の復元画像にこの変換モデルを適用し、得られた出力画像と元画像との間で RMSE（Root Mean Squared Error）を算出する。復元画像が一定の品質を持っていれば、元画像への変換は比較的正確に行えるため、RMSE は小さくなる。一方、復元画像がノイズを含み視認性が悪い場合、元画像への変換は困難になり、RMSE は大きくなると予測される。この特性を利用し、元画像と再劣化させた画像との RMSE が閾値を上回る画像を「復元失敗画像」と判定する。

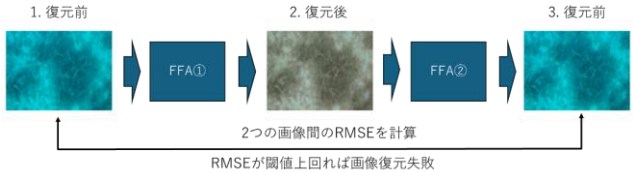

図 2 画像復元手法の概要図

3.2 疑似ラベル生成

提案手法では、2 段階の予測処理を通じて疑似ラベルを生成する。疑似ラベルの生成にはアノテーション画像を用いず、CLIP を用いたテキストプロンプトと画像内の各スーパーピクセルとの関係を用いて行う。以下に手法の詳細を示す。

3.2.1 第 1 段階：被度ゼロ画像の検出

第 1 段階の手順を図 3 に示す。第 1 段階では、藻場が全く存在しない極端な画像を検出する。本研究では、第 2 段階において SHAP[6]を用い、画像内のどの領域が CLIP の予測クラスに貢献したかを明らかにし、疑似ラベルの生成を行うが、藻場が存在しない画像に対して、シャープレイ値を算出すると、背景の視覚的特徴が藻場に類似すると判断され、誤って高い貢献度が算出される可能性がある。第 1 段階の処理でそのような画像を事前に除外することで、そのような誤認識の発生を防ぐことができる。第 1 段階の CLIP のテキストプロンプトは以下の 3 つである。

“a photo of no seagrass”, “a photo of without seagrass”, “a photo of some seagrass”

CLIP の分類結果が前者 2 つのいずれかであれば、その画像は被度ゼロとみなし、全ピクセルを背景とする疑似ラベルを生成する。一方で、“a photo of some seagrass”に分類された画像は、第 2 段階の処理で疑似ラベルを生成する。

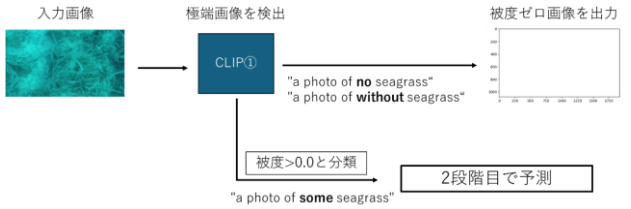

図 3 疑似ラベル生成 1 段階目の概要図

3.2.2 第 2 段階：シャープレイ値に基づく疑似ラベル生成

第 2 段階の手順を図 4 に示す。まず、SLIC アルゴリズムを適用し、入力画像からスーパーピクセルを生成する。その後、SHAP を用い、画像内のどの領域が CLIP の予測クラスに貢献したかを算出することで、画像内の藻場領域を推定する。第 2 段階の CLIP のテキストプロンプトを以下に示す。下記のテキストプロンプトについて、前半 5 個を「背景グループ」、後半 6 個を「藻場グループ」と呼ぶ。

背景グループ："a photo of sand", "a photo of water", "a photo of sand or water", "a blurry photo of water", "a blurry photo of sand",

藻場グループ："a blurry photo of seagrass", "a photo containing some seagrass","a photo of underwater plants", "a photo of underwater grass", “a photo of green, grass-like leaves“, a photo of seagrass"

各画像に対して CLIP と SHAP を適用し、各テキストプロンプトに対するシャープレイ値を算出する。各スーパーピクセルにおける藻場グループと背景グループのシャープレイ値の平均値を比較する。より高い平均値を持つグループに基づいて、各スーパーピクセルを「藻場」または「背景」として分類し、疑似ラベルを生成する。

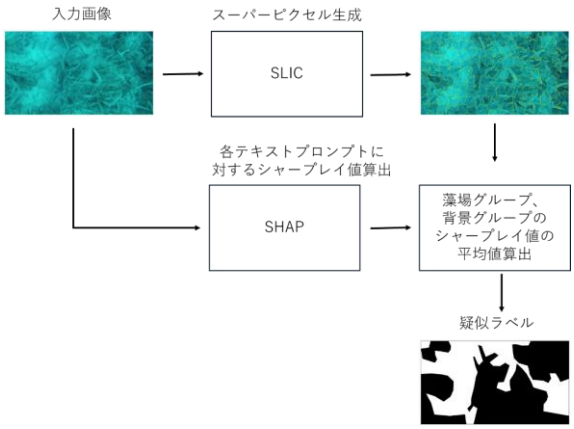

図 4 疑似ラベル生成 2 段階目の概要図

3.3 セマンティックセグメンテーション

セマンティックセグメンテーションを行うモデルとして SegFormer[7]を用いる。SegFormer は、Transformer ベースのセマンティックセグメンテーションモデルであり、エンコーダ部分とデコーダ部分で構成される。エンコーダ部分では、異なる解像度ごとで Transformer により特徴量を抽出する。デコーダ部分では、エンコーダから得られる異なる解像度の特徴マップを MLP Layer でアップサンプリングした後、MLP でセグメンテーションマップが生成される。SegFormer は既存手法よりも精度が高いことが報告されており、高精度な藻場推定が期待できる。

本研究では、復元成功画像を学習に用いたモデルと、復元失敗画像向けのモデルの２つを構築する。概要図を図 5 に示す。復元成功画像用のモデルでは、大量の復元成功画像と画像復元した少量のアノテーション付きデータを半教師あり学習を用いて SegFormer を学習する（図 5 のモデル①）。復元失敗画像用では、復元後の画像を用いず、大量

の復元前の元画像と復元前の少量のアノテーション付きデータを半教師あり学習を用いて SegFormer を学習する（図 5 のモデル②）。

推論時には、3.1 節で説明した RMSE に基づいて画像復元の成功・失敗を判定し、その結果に基づいてモデル①とモデル②のどちらを利用するかを決定する。

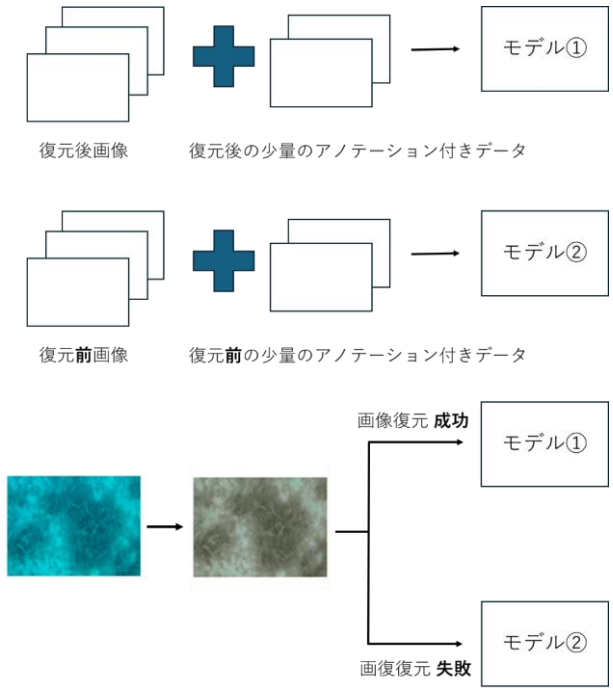

図 5 復元成功・失敗時におけるモデルの概要図

4. 評価実験

4.1 評価指標

評価指標には mean Intersection over Union （mean IU）, frequency weighted Intersection over Union （f.w. IU）, pixel accuracy(pixel acc.), mean accuracy(mean acc.)を用いる[8]。各指標の計算式は以下の通りである。

$$mean\ IU = \frac{1}{n_{cl}} \sum_i \frac{n_{ii}}{t_i + \sum_j n_{ji} - n_{ii}}$$

$$f.w.IU = \frac{1}{\sum_k t_k} \sum_i \frac{t_i n_{ii}}{t_i + \sum_j n_{ji} - n_{ii}}$$

$$pixel\ acc. = \frac{\sum_i n_{ii}}{\sum_i t_i}$$

$$mean\ acc. = \frac{1}{n_{cl}} \sum_i \frac{n_{ii}}{t_i}$$

n_{cl}はクラス数を、n_{ij}はクラス jに所属するピクセルがクラス iと予測されたピクセル数を表している。t_iは以下の計算式で、クラス iに所属する総ピクセル数を表している。

4.2 実験設定

FFA-Net モデルの損失関数として、MSE Loss と Perceptual Loss を組み合わせたものを設定した。MSE Loss の重みを 1.0, Perceptual Loss の重みを 0.1 として設定した。半教師あり学習の損失関数においては、アノテーションデータに対する損失の重みを 10.0、疑似ラベルに対する重みを 0.1 に設定した。また、セマンティックセグメンテーションモデルの学習に使用するアノテーションデータ付き画像の枚数は 3 枚とした。

4.3 画像復元

画像復元結果の例を図 6 に示す。また、画像復元が成功したかどうかの判定に用いる RMSE の閾値を 50 とした。

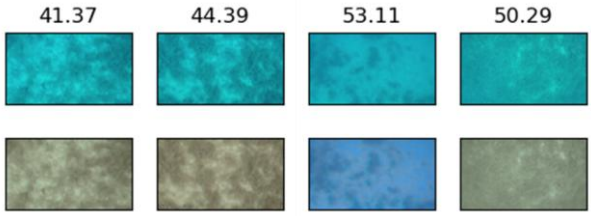

図 6　画像復元結果の例
（上から順に水中画像、復元画像、
数字は RMSE）

4.4 疑似ラベルの精度

生成された疑似ラベルの例を図 7 に示す。また、テストデータに対して、提案手法により生成された疑似ラベルの精度評価を表 1 に示す。

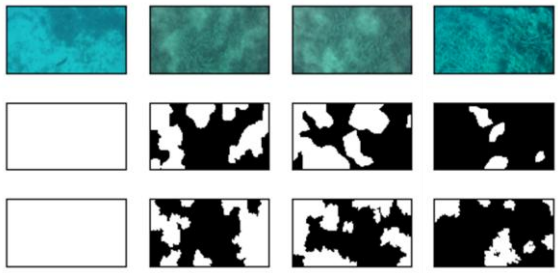

図 7　疑似ラベルの例
（上から順に水中画像、正解画像、
疑似ラベル、黒は藻場・白は背景を表す）

表 1　疑似ラベルの精度

mean IU	f.w. IU	pixel acc.	mean acc.
0.6937	0.7620	0.8090	0.7708

藻場の状態に応じた精度を評価するため、画像ごとの被度に基づき、被度が、0.1 未満の画像を「低被度藻場」、0.1 以上 0.9 以下の画像を「中被度藻場」、0.9 より大きい画像を「高被度藻場」と定義する。表 2 に各被度カテゴリごとでの精度を示す。

表 2　各被度ごとでの疑似ラベルの精度

被度	mean IU	f.w. IU	pixel acc.	mean acc.
低被度藻場	0.9496	0.9813	0.9844	0.9543
中被度藻場	0.3420	0.4169	0.5661	0.5315
高被度藻場	0.4671	0.6323	0.6561	0.5893

4.5 セマンティックセグメンテーション

提案手法の有効性を検証するため、教師あり学習を用い 3 枚のアノテーションデータ付き画像のみを学習したモデル（以下、少量の教師あり）と提案手法（以下、半教師あり）を比較する。さらに、それぞれにおいて画像復元（restoration）を適用した場合と適用しなかった場合についても評価を行った。各手法の精度指標を表 3 に示す。なお、先行研究は大量のアノテーションデータ付き画像を用いて学習されたモデルである。図 8 に提案手法（半教師あり w/ restoration）によるセマンティックセグメンテーション結果の例を、表 4 に各被度カテゴリごとでの精度を示す。表 3 より、画像復元を行わなかった半教師ありの手法が、全ての指標において他の手法よりも高い精度を示した。一方、画像復元を適用した場合は精度が低下し、画像復元の効果は確認されなかった。

表 3　少量の教師ありと提案手法と先行研究の精度

手法	mean IU	f.w. IU	pixel acc.	mean acc.
少量の教師あり w/o restoration	0.5704	0.6233	0.6662	0.6797
少量の教師あり w/ restoration	0.5147	0.5604	0.6015	0.6070
半教師あり w/o restoration	**0.7498**	**0.8503**	**0.8987**	**0.7993**
半教師あり w/ restoration	0.7393	0.8303	0.8769	0.7970
Weidmann	0.8778	0.9393	0.9642	0.9237

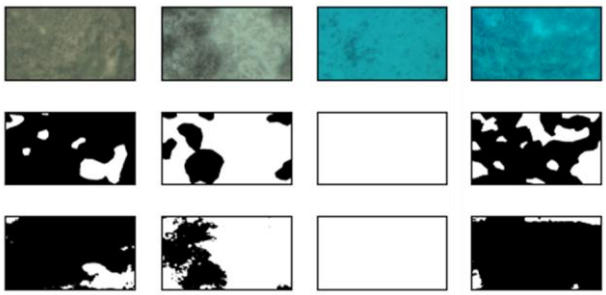

図 8　半教師あり w/ restoration による
セマンティックセグメンテーション結果の例
（上から順に入力画像、正解画像、予測画像）

表 4　各被度カテゴリごとでの精度
(半教師あり w/ restoration)

被度	mean IU	f.w. IU	pixel acc.	mean acc.
低被度藻場	0.9595	0.9919	0.9955	0.9615
中被度藻場	0.3926	0.4988	0.6471	0.5569
高被度藻場	0.6098	0.8492	0.8706	0.6718

5. おわりに

本研究では、画像復元と CLIP を活用した疑似ラベル生成による半教師あり学習手法を用いて、セマンティックセグメンテーションモデルの構築を試みた。その結果、画像復元の導入によって藻場推定の精度向上は確認できなかった。また、低被度藻場画像においては高い精度での推定が可能であり、高被度藻場画像に対してもある程度良好な精度を示した。しかし、中被度藻場画像では推定精度が十分ではなく、その要因としては疑似ラベルの精度が中被度藻場で低いことが挙げられる。今後の課題としては、中被度藻場の画像に対する疑似ラベル生成手法の改良および、推定精度の向上が挙げられる。

参考文献

[1] Gereon Reus et al. “Looking for seagrass: Deep learning for visual coverage estimation”, OCEANS, pp. 1-6, 2018.
[2] Franz Weidmann et al. “A Closer Look at Seagrass Meadows: Semantic Segmentation for Visual Coverage Estimation”, OCEANS, pp. 1-6, 2019.
[3] Radford Alec et al.“Learning transferable visual models from natural language supervision”, International conference on machine learning., pp.8748-8763, 2021..
[4] Raine, Scarlett, et al. "Image labels are all you need for coarse seagrass segmentation." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Visionm, 2024.
[5] M. J. Islam et al. "Fast Underwater Image Enhancement for Improved Visual Perception," in IEEE Robotics and Automation Letters, vol. 5, no. 2, pp. 3227 -234, 2020.
[6] Lundberg, Scott M., and Su-In Lee. "A Unified Approach to Interpreting Model Predictions." Advances in Neural Information Processing Systems 30, 2017.
[7] Xie, Enze, et al. "SegFormer: Simple and efficient design for semantic segmentation with transformers." *Advances in neural information processing systems* 34, 12077-12090, 2021
[8] Jonathan Long, Evan Shelhamer, Trevor Darrell. “Fully convolutional networks for semantic segmentation”, IEEE conference on computer vision and pattern recognition, pp. 3431-3440, 2015.