CCS に向けた深層学習による地質露頭画像のセグメンテーション

Segmentation of Geological Outcrop Images with Deep Learning Models for Carbon Capture and Storage

研究代表者　間所洋和　岩手県立大学ソフトウェア情報学部　教授
共同研究者　千代延俊　秋田大学大学院国際資源学研究科　教授
共同研究者　永吉武志　秋田県立大学生物資源科学部　准教授
共同研究者　ニックス ステファニー　岩手県立大学ソフトウェア情報学部　講師
Hirokazu Madokoro, Shun Chiyonobu, Takeshi Nagayoshi, and Stephanie Nix

Rapid climate change and global warming have widespread impacts on society, including ecosystems, water security, food production, health, and infrastructure. To achieve significant global emission reductions, approximately 74%, is expected to come from cutting carbon dioxide (CO2) emissions in energy supply and demand. Carbon Capture and Storage (CCS) has attained global recognition as a preeminent approach for the mitigation of atmospheric carbon dioxide levels, primarily by means of capturing and storing CO2 emissions originating from fossil fuel systems. Currently, geological models for storage location determination in CCS rely on limited sampling data from borehole surveys, which poses accuracy challenges. To tackle this challenge, our research project focuses on analyzing exposed rock formations, known as outcrops, with the goal of identifying the most effective backbone networks for classifying various strata types in outcrop images. We leverage deep learning-based outcrop semantic segmentation techniques using hybrid backbone networks to achieve accurate and efficient lithological classification, while considering texture features and without compromising computational efficiency. In the evaluation experiments conducted on ground-level images obtained using a stationary camera and aerial images captured using a drone, we successfully demonstrated the superior performance across all categories.

要旨

地球温暖化による気候変動が深刻化しており，既に生態系や水の安全保障と食料生産，健康と福祉，都市，居住地，インフラを含む人間システムに影響を及ぼしている．本研究では，二酸化炭素の回収・貯留（Carbon Capture and Storage：CCS）に注目し、深層学習に基づく意味的分類のセマンティックセグメンテーションを用いた露頭画像の地層分類のために最適なバックボーンの探求を目的とした．ドローンによる撮影した13枚の露頭画像に対して，データセットの拡張と評価実験を行った．実験結果より，ViT（Vision Transformer）に対してCNN（Convolutional Neural Network）が優位性を示すことを明らかにした．また，画素占有率が偏っている場合，精度に影響を及ぼすことを明らかにした．更に，各クラスの特徴を捉え，未知のデータに対して対応できることを明らかにした[1]．

1．　まえがき

地球温暖化による気候変動が深刻化しており，既に生態系や水の安全保障と食料生産，健康と福祉，都市，居住地，インフラを含む人間システムに影響を及ぼしている．CO2 排出がゼロに達する時点で世界全体の排出削減量のうち約 74%がエネルギー供給・需要における CO2 排出量削減によって達成されると想定される．主流の方法としては，再生可能エネルギが挙げられるが，需要を満たすためには化石燃料システムも組み合わせる必要がある．そこで本研究では，化石燃料システムからの炭素回収率が 90～95%強と想定される二酸化炭素回収・貯留（Carbon Capture and Storage：CCS）に注目した．

図1 対象とした地質露頭（秋田県男鹿市船川）

本研究では，図1に示す表土や植物などによって覆われている地層が地表に現れた部分である露頭に着目した．撮影した露頭画像から地層の分布を明らかにし，高精度な地層モデルを作成することで貯留に最適な場所の同定を目指した．そこで本研究では，深層学習（Deep Learning）に基づく意味的分類（セマンティックセグメンテーション）を用いた露頭画像の地層分類のために最適なバックボーンの探求を目的とする．

2．関連研究

セマンティックセグメンテーションは，画像内の対象物を画素単位で分類する手法である．コンピュータビジョンにおける基本的な技術であり，自動運転などの様々な実用的なタスクに活用されている．これまでのセマンティックセグメンテーションのアプローチとしては，CNN（Convolutional Neural Networks）[1]が主流であった[2]．しかしながら，2020年にDosovitskiyら[3]がViT(Vision Transformer)を発表し，最新手法であったBiT(Big Transfer)[4]を超える精度を達成したことから，ViTを用いた研究が活発化している．ViTは自然言語処理において主流を築いたTransformer[5]を画像認識に適応したアーキテクチャである．

Transformerは，CNNの近い位置の情報は関係が深いというデータが持つ固有の構造(帰納バイアス)を持たないため，汎化のために大量の訓練データを必要とする[vit]．Geirhosら[6]は，CNNの分類特性が人間の知覚とは異なり，物体形状よりもテクスチャを重視していることを明らかにした．一方で，Tuliら[7]は，ViTの分類特性が物体形状に偏りがあり，人間の知覚に近いことを明らかにした．近年，CNNやViT，CNN+ViT型のバックボーンに加えて，多層パーセプトロン(Multi-Layer Perceptron: MLP)だけのシンプルなバックボーン[8]が注目されている．分類タスクでは，BiTやViTに匹敵する性能を示す一方で，セグメンテーションには応用されていない．

地学分野では，従来，高精度の地上レーザスキャナ測量によって3Dモデルを作成してきた[9]．しかしながら，この方法には，測量機器重量，複数のフィールドベースの位置からのスキャンの必要性，取得時間の長さなどの制限がある．このため，地上からデータが取得できない場所や危険な場所では，モデリングが困難になる．そのような場所では，カメラを搭載したドローンが使用されている．

Corradettiら[10]は，ドローンを使用して，ほぼ垂直な崖で構成された露頭を撮影した写真から作成した3Dモデルを使用して，露頭の亀裂を解析し，その広がり方を明らかにしている．

Sharadら[11]は，地形が複雑で，測定が困難かつ危険な地すべりの高解像度画像を収集するためにドローンを使用して，撮影画像から計測精度がcmレベルの3Dモデルを作成した．

Javierら[12]は，スペイン北西部にある古代ローマ時代の金鉱跡の同定と解釈のために，ドローンを用いて高精度で高解像度の3Dモデルを作成し，発掘場や運河，貯水池，排水溝などの考古学的記録を明らかにした．このように，地学分野におけるドローンを使用した研究は，活発化している．しかしながら，その多くが撮影した写真から作成した3Dモデルを使用した地形の解析であり，露頭に対するセマンティックセグメンテーションに基づく地層分類は，報告されていない．

3．提案手法

複数の深層学習モデルを融合した提案手法の全体構成を図2に示す．

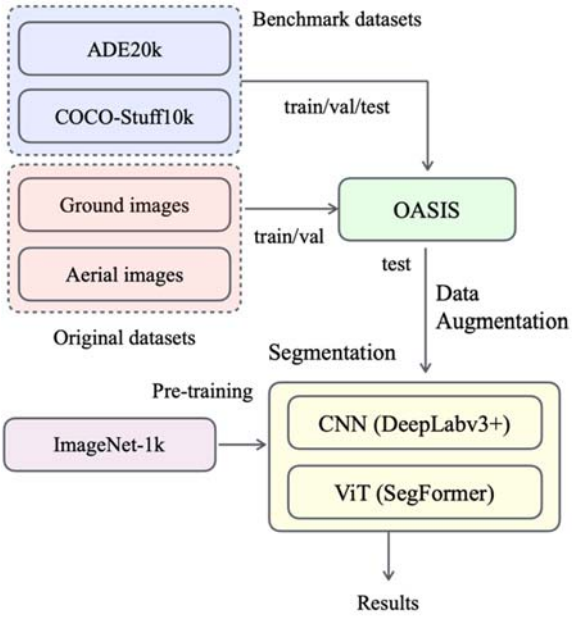

図2 提案手法の構成

K-Net [13] は汎用性が高く， Kernel Update Head を追加するだけでどのような手法にも適用できる．Kernel Update Head の処理を繰り返すことで，ノイズが軽減されカーネルとマスクの予測値が徐々に洗練される．SegFormer [14] は ViT(Vision Transformer) ベースの手法である．エンコーダ部分に階層型 Transformer，デコーダ部分に All-MLP を採用することで高解像度の細かい特徴と低解像度の粗い特徴の情報を集約し，豊富な特徴量を得る．また，画像生成に GAN (Generative Adversarial Network) に基づく生成モデルである OASIS (Only Adversarial Supervision for Semantic Image Synthesis) [15] を用いる．この手法は，クラスバランシングを組み込んでいるため，クラスの不均衡による精度低下を軽減できる．

セマンティックセグメンテーションの評価指標には，mIoU(mean Intersection over Union)を使用する．IoU は，予測領域と正解領域とが重なる程度の交差度を表し，mIoU はすべてのクラスの平均 IoU を示す．IoU は以下の式で求められる．

$$IoU = \frac{TP}{TP + FP + FN} \qquad (1)$$

ここで，予測とクラスがともに真の場合は真陽性(TP)，偽の予測に対してラベルが真の場合は偽陽性(FP)，真の予測に対してラベルが偽である場合は偽陰性(FN)を表す．OASIS の評価指標には，フレシェ開始距離 (Fréchet Inception Distance：FID)を使用する．FID は，各画像の特徴ベクトルから，各特徴量に対する距離を算出し，原画像と生成された画像の類似性を定量的に評価している．ここで，生成されたデータと入力されたデータから得られた特徴ベクトルの大きさをそれぞれ m_w, m と定義する．

$$FID = \|m - m_w\|^2 + \mathrm{Tr}\left(C + C_w - 2\sqrt{CC_w}\right) \qquad (2)$$

ここで各々の特徴ベクトルの共分散行列をそれぞれ C_w, C と定義する．

4．予備実験

4．1　データセット

本実験では，13 枚の露頭画像に対して，地層を専門家が振り分けた black，red，cyan，yellow の 4 クラスに加えて，アノテーションされていない画素に green を割り当てた 5 クラスのデータセットを使用した．しかし，green には影など地層以外の要素が含まれているため，green を計算から除外し，結果として，4 クラスで分類を行った．

入力には，図 3 に示す 13 枚の各画像から 256×256 画素でランダムサンプリングした特徴量を使用する．サンプリング数は 64，128，256，512 枚とした．表 1 に各サンプリング数におけるそれぞれのクラスの画像全体に占める画素の割合(以下，画素占有率)を示す．学習データとテストデータは 13 枚の各画像からサンプリング数ずつランダムサンプリングを行った後，無作為に 9:1 の比率で分割した．

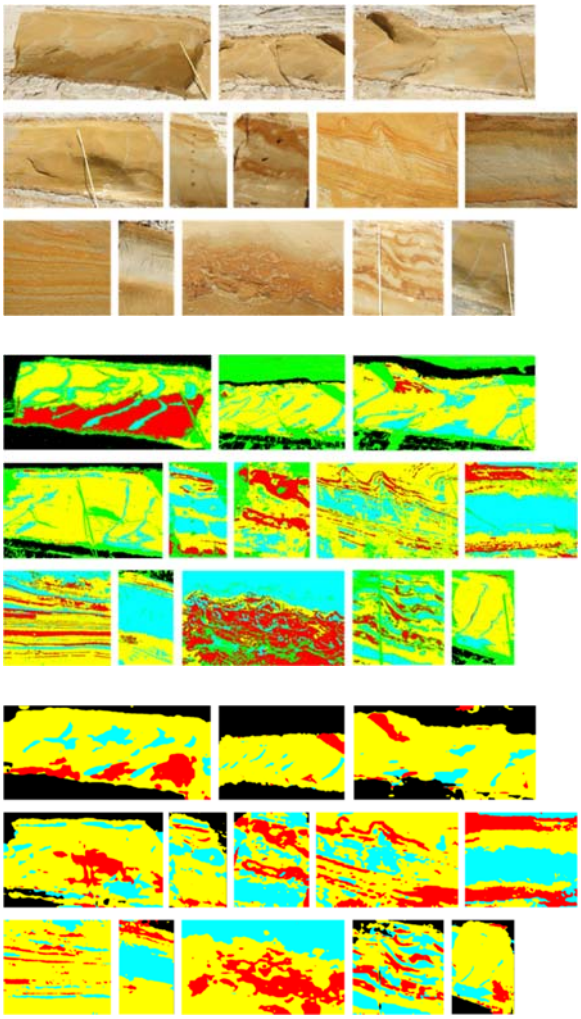

図 3 対象画像（上段），GT（中段），分類例（下段）

4．2　実験結果

本データセットにおける代表的な手法の精度比較を行った．手法としては CNN ベースのものとして DeepLabv3+，K-Net+DeepLabv3，OCRNet，U-Net を使

用した．また，ViT ベースの手法として K-Net+Swin Transformer, SegFormer, SETR を使用した．そして，CNN と ViT を組み合わせた手法として Twins を使用した．図 4 に結果を示す．結果は，CNN が ViT に対して優位性を示した．また，K-Net (DLabv3) と SegFormer の mIoU はそれぞれ 93.18%, 93.06%と特に高い精度を示した．

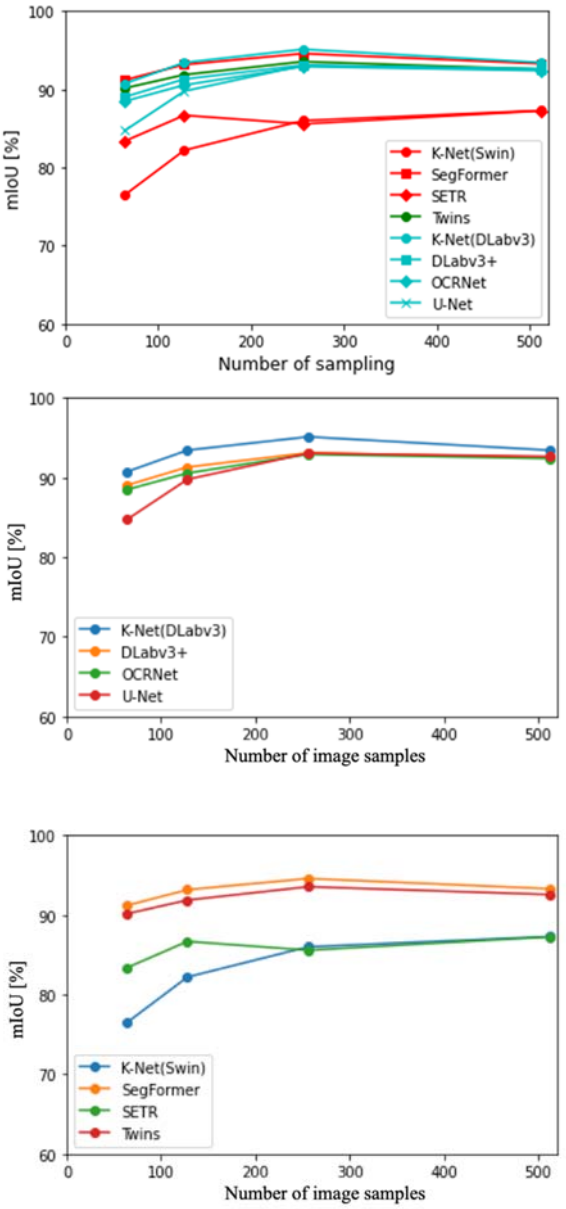

図 4 バックボーン別の全実験結果（上段），CNN 系バックボーンの結果（中段），ViT 系バックボーンの結果（下段）

続いて、13 枚の各画像に対して推論を行った．サンプリング数は 256 枚とし，推論対象の画像をテストデータ，残りの 12 枚の画像を学習データとした．手法は，K-Net (DLabv3) と SegFormer を使用した．全画像の平均 mIoU は，K-Net で 42.66%，SegFormer で 50.84%を示した．このことから，データが少ない場合，ViT に優位性が見られると考えられる．また，クラスに偏りがある場合，精度に影響を与えることを明らかにした．

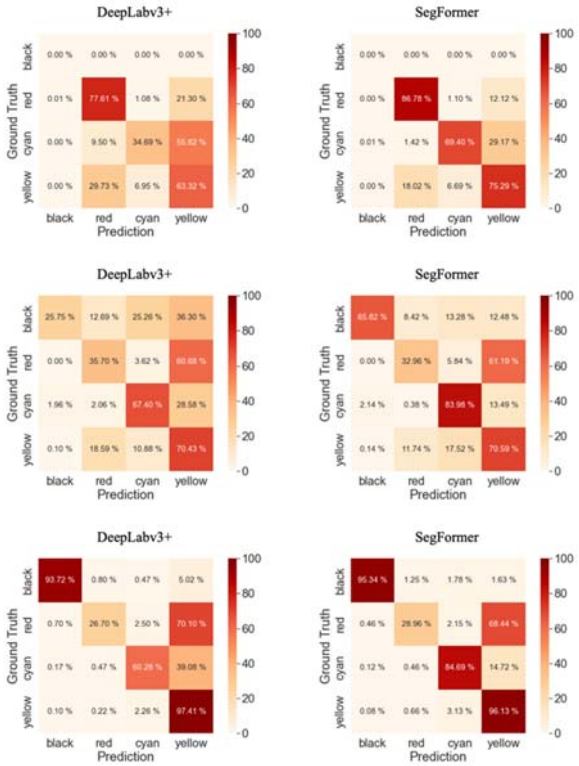

図 5 データ拡張前（上段），10000 回学習後（中段），20000 回学習後（下段）の各結果における混同対称マトリクス

データセットを拡張するために OASIS を使用し，画像生成を行った．実験には，サンプリング数 256 枚のデータセットを使用した．データ拡張前，10000 回学習後，20000 回学習後の各結果における混同対称マトリクスを図 5 に示す．学習回数が 20000 回のときに FID の値は 363.34 を示し，最も低い値を示した．また，生成した画像を用いてデータセットを拡張し，評価実験を行った．データセットはサンプリング数 256 枚の画像に OASIS で生成した 333 枚の画像を加えた計 3661 枚の画像を使用した．手法は，K-Net (DLabv3) と SegFormer を使用した．K-Net と SegFormer の mIoU はそれぞれ 95.02%と 95.51%であり，データセット拡張前と比べて，それぞれ 2.33%，1.96%上昇していた．したがって，OASIS によるデー

タセット拡張が精度向上に寄与すると考えられる．

５．公開ベンチマークによる評価実験

公開ベンチマークデータセットを用いて，ViT と CNN のクラスごとの分類特性の評価と精度比較を行った．ベンチマークデータセットとして COCO-Stuff 10K [16] と ADE20K [17] を使用した．手法は，K-Net (DLabv3) と SegFormer を使用した．

５．１　COCO-Stuff 10K 分類結果

COCO-Stuff 10K ベンチマークは，車や人などの形状の輪郭が明確な物体を表す things クラスが 80 クラス，草、空などの定型の輪郭を持たない背景領域を表す stuff クラスが 91 クラスの計 171 クラスで構成される．学習画像は 9000 枚，テスト画像は 1000 枚である．各バックボーンにおける分類結果の比較分布を図 6 に示す． K-Net と SegFormer の mIoU はそれぞれ 29.36%と 41.38%であった．クラス全体では 171 クラス中 154 クラス(90%)で SegFormer に優位性が示された．things クラスでは 80 クラス中 4 クラス(5%)，stuff クラスでは 91 クラス中 13 クラス(14%)で K-Net に優位性が見られた．

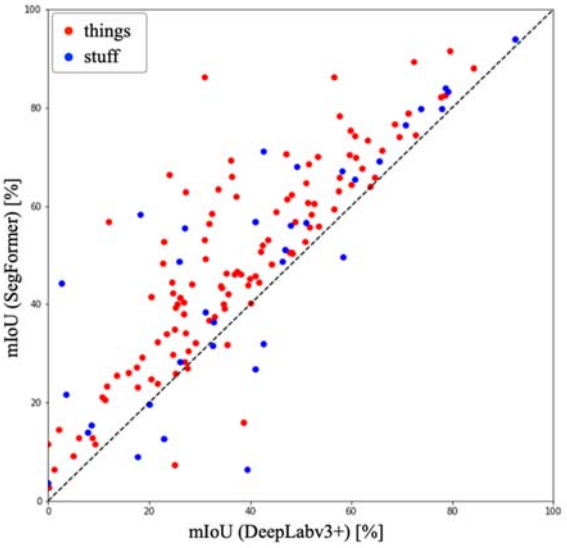

図 6 各バックボーンにおける比較結果 (COCO-Stuff 10K)

５．２　ADE20K 分類結果

ADE20K ベンチマークは，25574 枚の学習データと 2000 枚のテストデータから構成される．データセットの内訳は，things が 115 クラス，stuff が 35 クラスである．各バックボーンにおける分類結果の比較分布を図 7 に示す．K-Net と SegFormer の mIoU はそれぞれ 38.78%と 48.40%であった．クラス全体では 150 クラス中 138 クラス(92%)で SegFormer に優位性が示された．things クラスでは 115 クラス中 4 クラス(3%)，stuff クラスでは 35 クラス中 8 クラス(23%)で K-Net に優位性が示唆された．

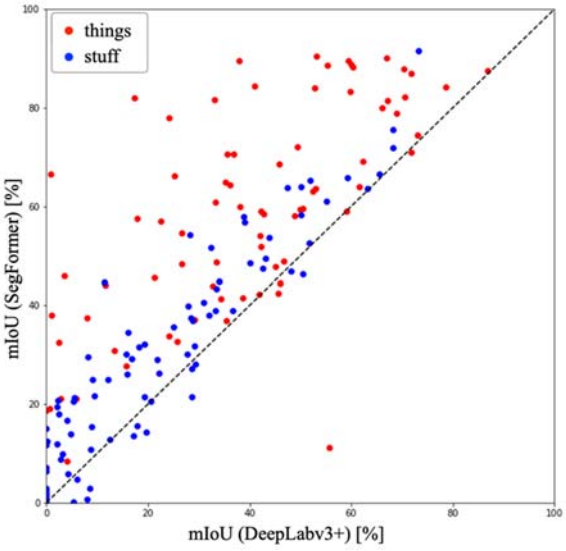

図 7 各バックボーンにおける比較結果 (ADE 20K)

５．３　考察

いずれのデータセットにおいても ViT(SegFormer)が CNN(K-Net)より高い精度を示した．K-Net に優位性が見られたクラスの割合は，stuff クラスの方が高かった．これは，分類がテクスチャに依存するためであると考える．

６．応用実験

６．１　データセット

本実験では空撮画像に対して，セマンティックセグメンテーションを行った．空撮の様子を図 8 に示す．ドローンは DJI 社製の Mavic 2 Pro を使用した．

図 8 空撮の様子

空撮により取得した画像と GT 画像を図 9 に示す。解像度が 5464×3640 画素の画像に対して，専門家がオリジナルデータセットと同様にアノテーションを施したデータを使用した．推論は K-Net (DeepLabv3) と SegFormer の OASIS による拡張データセットを用いて学習したモデルを使用した．また，推論時に空撮画像を縦横 8 等分にして入力した．すなわち，入力サイズは683×455 画素，データ数は64 枚である．

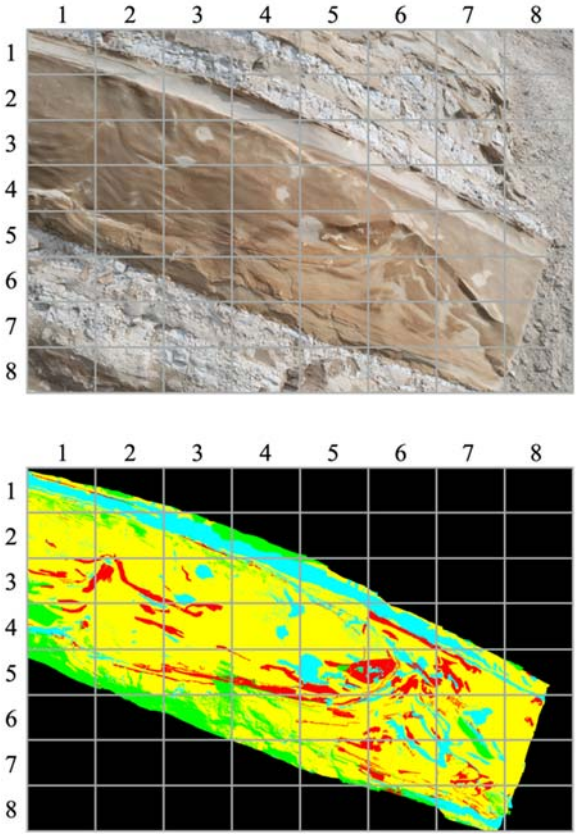

図 9 空撮画像（上段），GT 画像（下段）

6．2　実験結果

本実験における分類精度として，mIoU は K-Net で 24．17%，SegFormer で 40．30%を示し，いずれのクラスにおいても 50%を下回った．これは，データの多様性の不足が原因であると考える．画像毎では，mIoU が上位 5 枚の画像は中心部に集中しており，最も mIoU が高い画像は SegFormer で 49．03%を示した．したがって，オリジナルデータセットから傾向を捉え，未知の画像に対応できていると考えられる．一方で，下位 5 枚の画像は，すべて外側の black のみが現れる部分であった．

そこで，一部の画像に専門家が再度アノテーションを施した．再アノテーション前後の分類結果を図 10 に示す。この画像では，mIoU は K-Net で 10．75%，SegFormer で 50．09%を示し，再アノテーション前と比べて，それぞれ 10．23%，39．69%上昇した．これは，深層学習が人間にアノテーションの再考を提案できる可能性を示唆している．

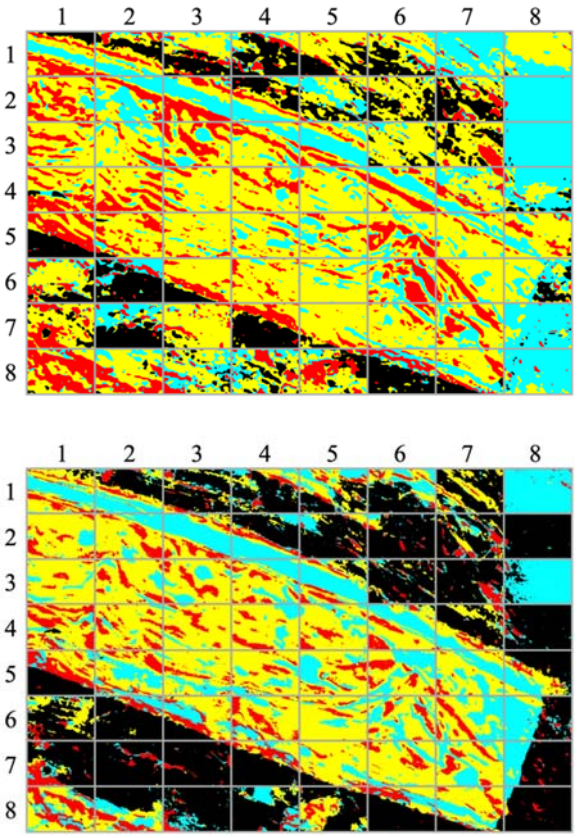

図 10 分類結果（上段），再アノテーション後の分類結果（下段）

7．まとめ

地質露頭における地層分類を対象とした本研究では，形状が曖昧な非定型の対象は分類がテクスチャに依存するため，CNN に優位性が見られることを明らかにした[1]．また，データが少ない場合，ViT に優位性が見られることを明らかにした．さらに画素占有率の偏りが精度に影響を及ぼすことを明らかにした．OASIS で生成した画像でデータセットを拡張することで精度向上に寄与することを明らかにした．空撮画像に対して推論を行い，13 枚のデータから，

各クラスの特徴を捉え，未知のデータに対して対応できることを明らかにした.

今後の課題としては,データの多様性の向上やGANによる画像生成のテクスチャや色の再現性の向上,推論結果を元にしたアノテーション修正の提案が挙げられる.

発表論文

[1] Hirokazu Madokoro, Kodai Sato, Stephanie Nix, Shun Chiyonobu, Takeshi Nagayoshi, and Kazuhito Sato "OutcropHyBNet: Hybrid Backbone Networks with Data Augmentation for Accurate Stratum Semantic Segmentation of Monocular Outcrop Images in Carbon Capture and Storage Applications," Sensors (Special Issue: Machine Learning Based Remote Sensing Image Classification), vol.23, no.8809, 2023. (doi:10.3390/s23218809).

参考文献

（1） Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton, "Imagenet classification with deep convolutional neural networks.", Advances in Neural Information Processing Systems (NeurIPS), 2012.

（2） Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, Demetri Terzopoulos, "Image Segmentation Using Deep Learning: A Survey", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.44, Issue.7, pp.3523-3542, 2021.

（3） DosoViTskiy Alexey, Beyer Lucas, Kolesnikov Alexander, Weissenborn Dirk, Zhai Xiaohua, Unterthiner Thomas, Dehghani Mostafa, Minderer Matthias, Heigold Georg, Gelly Sylvain, Uszkoreit Jakob, Houlsby Neil, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale", arXiv:2010.11929, 2020.

（4） Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, Neil Houlsby, "Big Transfer (BiT): General Visual Representation Learning", arXiv:1912.11370, 2020.

（5） Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin, "Attention is All you Need", NeurIPS, 2017.

（6） Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, Wieland Brendel, "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness", arXiv:1811.12231, 2018.

（7） Shikhar Tuli, Ishita Dasgupta, Erin Grant, Thomas L. Griffiths, "Are Convolutional Neural Networks or Transformers more like human vision?", arXiv:2105.07197, 2021.

（8） Ilya Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung,Reas Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, Alexey Dosovitskiy, "MLP-Mixer: An all-MLP Architecture for Vision", arXiv:2105.01601, 2021.

（9） Rémy Richet, Jean Borgomano, Erwin W. Adams, Jean-Pierre Masse, Sophie Viseur, "Numerical Outcrop Geology Applied to Stratigraphical Modeling of Ancient Carbonate Platforms: The Lower Cretaceous Vercors Carbonate Platform (Se France)", In:Ole J. Martinsen, Andrew J. Pulham, Peter D.W. Haughton, Morgan D. Sullivan, "Outcrops Revitalized: Tools, Techniques and Applications", SEPM Concepts in Sedimentology and Paleontology, pp.195-210, 2011.

（10） A. Corradetti, S. Tavani, M. Parente, A. Iannace, F. Vinci, C. Pirmez b, S. Torrieri, M. Giorgioni, A. Pignalosa, S. Mazzoli, "Distribution and arrest of vertical through-going joints in a seismicscale carbonate platform exposure (Sorrento peninsula, Italy): insights from integrating field survey and digital outcrop model", Journal of Structural Geology 108 (2018) pp.121-136.

（11） Sharad Kumar Gupta, Dericks P. Shukla, "3D Reconstruction of a Landslide by Application of UAV ¥& Structurefrom Motion", AGILE 2017 - Wageningen, May pp.9—12, 2017,

（12） Javier Fernández-Lozano, Gabriel Gutiérrez-Alonso, "Improving archaeological prospection using localized UAVs assisted photogrammetry: An example from the Roman Gold District of the Eria River Valley (NW Spain)", Journal of Archaeological Science: Reports 5 (2016) pp.509-520.

（13） Wenwei Zhang, Jiangmiao Pang, Kai Chen, Chen Change Loy, "K-Net: Towards Unified Image Segmentation", NeurIPS, 2021.

（14） Xie Enze, Wang Wenhai, Yu Zhiding, Anandkumar Anima, Alvarez Jose M, Luo Ping, "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers", arXiv:2105.15203, 2021.

（15） Vadim Sushko, Edgar Schönfeld, Dan Zhang, Juergen Gall, Bernt Schiele, Anna Khoreva, ``OASIS: Only Adversarial Supervision for Semantic Image Synthesis", International Journal of Computer Vision, vol.130, pp.2903—2923, 2022.

（16） H. Caesar, J. Uijlings, V. Ferrari," COCO-Stuff: Thing and Stuff Classes in Context", In Computer Vision and Pattern Recognition (CVPR), 2018.

（17） Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., & Torralba, A.," Semantic understanding of scenes through the ade20k dataset.", International Journal of Computer Vision, 127(3), 302-321,2019.