季節のコントラスト：未キュレーションのリモートセンシングデータからの教師なし事前学習

Oscar Man˜as[1,2] Alexandre Lacoste[1] Xavier Giro´i-i-Nieto[2] David Vazquez[1] Pau Rodriguez[1]
[1]エレメントAI [2]カタルーニャポリテニカ大学
oscmansan@gmail.com, pau.rodriguez@servicenow.com

要旨

リモートセンシングと自動地球モニタリングは、防災、土地利用モニタリング、気候変動への取り組みなど、地球規模の課題を解決する鍵である。膨大な量のリモートセンシングデータが存在するが、そのほとんどはラベル付けされていないため、教師あり学習アルゴリズムにはアクセスできない。伝達学習アプローチは、深層学習アルゴリズムのデータ要件を削減することができる。しかし、これらの手法のほとんどは、Ima- geNet上で事前に訓練されており、ドメインギャップのため、リモートセンシング画像への一般化は保証されていない。本研究では、リモートセンシング表現のドメイン内事前学習のためにラベルなしデータを活用する効果的なパイプラインであるSeasonal Contrast (SeCo) を提案する。SeCoパイプラインは2つの部分から構成される。第一に、異なるタイムスタンプで複数の地球上の位置からの画像を含む、ラベル付けされていない大規模なリモートセンシングデータセットを収集する原理的な手順である。第二に、時間と位置の不変性を利用し、リモートセンシング応用のための転送可能な表現を学習する自己教師付きアルゴリズムである。SeCoで学習されたモデルは、ImageNetで事前に学習されたモデルや、最先端の自己教師付き学習手法よりも、複数の下流タスクにおいて優れた性能を達成することを実証的に示す。SeCoのデータセットとモデルは、移行学習を促進し、リモートセンシング応用における急速な進歩を可能にするために公開される予定である。[1]

1. はじめに

リモートセンシングは、土地利用監視 [12]、事前判断型農業 [29]、災害防止 [37]、山火事検出 [11]、媒介感染症監視 [20]、気候変動への取り組み [33] など、多くの用途で重要性を増している。最近のディープラーニングとコンピュータビジョンの進歩も相まって、このような分野への応用はますます広がっている。

[1] コード、データセット、訓練済みモデルは https://github.com/ElementAI/seasonal-contrast で入手できる。

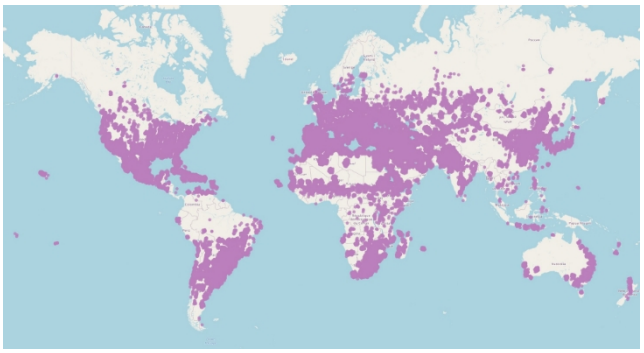

図1.**Seasonal Contrast (SeCo)データセットの分布。**各ポイントはサンプリングされた場所を表す。海や砂漠のような単調な場所を避けるため、画像は人里周辺で収集されている。

リモートセンシングやその他の地理空間データのストリームを自動解析することによって、地球規模の問題を監視するための大きな可能性がある。

リモートセンシングは、膨大な量のデータを提供する。地球観測衛星の数は増加の一途をたどっており、現在700以上の衛星が軌道上にあり、毎日テラバイトの画像データを生成している[30]。しかし、関心のある多くの下流タスクは、アノテーションの不足に制約されている。アノテーションは、専門家の知識や高価な地上センサーを必要とすることが多いため、取得には特にコストがかかる。近年、ラベル付けされたデータの必要性を軽減するために、多くの技法が開発されている[24, 26, 25]が、リモートセンシング画像への応用はほとんど未開拓である。

さらに、既存のリモートセンシング・データセット[38, 19, 42]は、バランスのとれた多様なクラスを形成するために、高度にキュレーションされている。単にラベルを捨てるだけでは、このような注意深い事例の選択を取り消すことはできない。我々の目標は、一般に公開されている膨大な量のリモートセンシングデータを利用し、真の教師なし方法で優れた視覚表現を学習することである。これを可能にするために、我々はSentinel-2 [10]のタイルからリモートセンシングデータセットを構築する。

リモートセンシングデータに特有なもう一つの特徴は、衛星の再 現時間であり、これは、地表の同じ地点の画像を繰り返し撮影するシステムの能力を表す。Sentinel[10]やLandsat[35]のような公的資金による衛星コンステレーションでは、再訪問時間は数日のオーダーである。この時間的な次元は、画像の人工的な補強を補完する自然変動の追加的なソースを提供する。例えば、雪山の頂上が雪解け時にどのように見えるか、あるいは作物の様々なステージが季節を通してどのように変化するかを示すことは、いくら人工的に補強してもできない。

自己教師付き学習法は、ラベル付けされていない膨大な量のデータから学習するための効果的な方法論として最近登場した。対照学習法は、意味的に類似した画像（すなわち、正対する画像）の表現を押し合う。ラベルが利用できないため、自然画像を扱う従来の対比学習法では、同じ画像（ビュー）の異なる人工的な拡張を正対として使用する。リモートセンシング画像の場合、我々は時間情報を活用し、異なる時点の同じ場所からの画像のペアを得ることを提案する。我々は、海水の変化は人工的な変換よりも意味的に意味のあるコンテントを提供し、リモートセンシング画像はこの自然な補強を無料で提供すると主張する。

我々は、リモートセンシングのアプリケーションのために、豊富で転送可能な表現を事前に学習するための新しい方法論であるSeasonal Contrast (SeCo)を提案する。SeCoは、教師なしデータ取得手順と、取得データから学習するための自己教師付き学習モデルの2つの部分から構成される。自己教師付き学習モデルは、季節変化に対して不変な表現を奨励することは、強い帰納的バイアスであるという観察に基づいて設計されている。この性質は、季節変動によって予測が変化しない特定の下流タスク（例えば、土地被覆分類、農業パターン区分、建物検出）には有益であるが、季節変動が重要である下流タスク（例えば、デフォールステーション追跡、変化検出）には有害である。我々は、リモートセンシング画像が適用可能な下流タスクに適した表現を学習したい。

視覚表現が常に時間に対して不変であるように制限することなく、時間情報を活用するために、我々は複数の埋め込み部分空間[47]のアイデアを使用する。すべての拡張に対して不変な単一の埋め込み空間に画像をマッピングする代わりに、個別の埋め込み部分空間を構築し、季節的変化に対して可変または不変になるように最適化する。我々は、異なる分散と不変を符号化する共通の表現を生成する共有バックボーンを持つマルチヘッドアーキテクチャを使用する。一旦モデルが学習されれば、この表現は幅広いリモートセンシングの下流タスクに適用できる、

ここでモデルは、表現に取り込まれたさまざまな変動要因を選択的に利用することができる。

いくつかのリモートセンシングデータセットとタスクで SeCo を評価する。BigEarthNet[38] と EuroSAT[19] を用いた土地被覆分類、および OSCD[8]を用いた変化検出の実験により、SeCoの事前学習が一般的なIma- geNet[36]やMoCo[18]の事前学習よりもリモートセンシングタスクに対して効果的であることが実証された。

まとめると、我々の貢献は以下の通りである:

- リモートセンシング画像からラベル付けされていないデータセットを収集する一般的な方法について述べる。この方法を用いて、人間の監視なしにSentinel-2のタイルからリモートセンシングデータセットを構築する。
- 我々は、最近の対照的な自己教師付き学習法と衛星から得られる時間的情報を組み合わせることで、季節の変化に対して不変で、同時に変化する優れた視覚表現を学習する。
- BigEarthNetとEuroSATの土地被覆分類、およびOSCDの変化検出において最先端の結果を得た。

2. 背景

自己教師あり学習は教師なし学習の一分野であり、データそのものが監視を行う。主なアイデアは、データの一部を隠蔽または摂動させ、可視データからそれを予測することをネットワークに課すことである。このタスクはプレテキストタスク（またはプロキシロス）と呼ばれ、ネットワークはこのタスクを解決するために、データについて気になること（例えば意味的な表現）を学習することを余儀なくされる。例えば、パッチの相対位置の予測[9]、ジグソーパズルの解法[31]、回転の予測[13]、色付け[48]などである。

最近では、対照的な口実タスク [46, 32, 41, 18、28, 5, 16, 4]は、様々な下流タスクで優れた性能を示し、自己教師あり学習のサブフィールドを支配してきた。直観的に言えば、対照学習法は、類似する例の表現を引き合わせる一方で、非類似の例の表現を引き離す。例はラベル付けされていないので、これらの方法は、各例がそれ自身のクラスを定義し、それに属すると仮定する。従って、正のペアは同じ例にランダムな追加を適用することで生成され、負のペアはデータセットの他の例から得られる。

ここで、与えられた例xは、2つのビュー、クエリx^q とキーx^k に拡張され、エンコーダネットワークfは、埋め込み空間に例をマップし、クエリ$q = f(x^q)$ の再提示は、その指定されたキー$k^+ = f(x^k)$ の表現に、そのクエリから来る任意の否定的なキーk^- の表現よりも近くなければならない。

この目的のために、正負のペアのバッチに対して対比的な目的が最適化される。一般的な選択は、InfoNCE損失[32]である：

$$L = -\log \frac{\exp(q \cdot k^{+}/\tau)}{\exp(q \cdot k^{+}/\tau) + \sum_{k^{-}} \exp(q \cdot k^{-}/\tau)} \quad (1)$$

ここで、τは距離分布をスケーリングする温度ハイパーパラメーターである。

3. 方法

この方法は、教師なし事前学習データセットを収集する一般的な手順（セクション3.1）と、このデータを活用するための自己教師付き学習方法（セクション3.2）から構成される。

3.1. 教師なしデータセットの収集

リモート・センシングは膨大な量の画像データを提供するが、注釈は通常乏しく、専門知識や地上センサーが必要になることが多い[21]。大量の衛星画像で学習するために、我々はSentinel-2 [10]パッチの新しいデータセットを収集する。Sentinel-2の画像は、10m、20m、60mの解像度の12スペクトルバンド（RGBとNIRを含む）で構成され、再訪問時間は約5日である。Google Earth Engine [15]を使用して、Sentinel-2の画像パッチを処理し、ダウンロードする。

各パッチはおよそ2.65×2.65kmの領域をカバーしている。各地点で、異なる日付の画像を5枚ずつダウンロードする。

これは、その地域で1年間に起こった季節的な変化をとらえたものである。同じ時期の画像を取得するのを避けるため、各ロケーションで日付を最大1年間ジッターさせた。また、雲の割合が10%以上のSentinel-2のタイルをフィルタリングした。合計で約100万枚のマルチスペクトル画像パッチが得られ、これは合計3870億画素以上に相当する。

サンプリング戦略 我々の目的は、様々な下流タスクに使用できるエンコーダを学習することである。そのためには、地球上の様々な地域からサンプリングする必要がある。一様なサンプリングでは、画像の種類に大量の冗長性が生じる。例えば、海は地球の71%、森林は31%、砂漠は33%を占めている。これを回避するために、変動性の大部分は都市周辺のより広い地域で観測されると仮定する。都市自体にはさまざまな建築物があり、都市から数キロメートル離れた場所では、さまざまな作物や工業施設がよく観察される。最後に、都市から50km～100kmの範囲では、通常、自然環境が観察される。したがって、このヒューリスティックに従って、都市周辺のサンプリングを行った（図1の結果を参照）：

1. 人口が最も多い10kの都市から一様にサンプリングし、都市の中心を中心に標準偏差50kmに及ぶガウス分布から座標のセットをサンプリングする。
2. 過去1年間の基準日をランダムに選択する。サンプリング日を得るために、3ヶ月単位で定期的に追加する。
3. 各日付を中心とした15日間の範囲について、座標と交差する雲量が10%未満のSentinel-2タイルが存在するかどうかをチェックする。
4. すべての日付にこの場所の有効なセンチネル-2タイルが存在する場合、すべての画像パッチを処理し、ダウンロードする。そうでなければ、ステップ1に進む。

得られた画像が多様で、情報量が多く、雲がないことを確認するために、追加のデータクリーニングは行わない。我々のデータセットは自動で構築されるため、より多くのデータ（より多くの場所、場所ごとのより多くの日付）を簡単に収集することができる。しかし、本研究では、Im- ageNet [36]と比較できるようにするため、規模を合計1M画像に制限している。

3.2. 季節のコントラスト

時間情報を持つリモートセンシング画像の教師なしデータセットが与えられたとき、データの構造を利用した表現を学習する。我々は[47]からヒントを得て、マルチオーグメンテーションコントラスト学習法を開発する。このアプローチは、人為的な補強によって発生する形成内損失を選択的に防ぎ、リモートセンシング画像の季節変化によって提供される自然な補強でそれを拡張することができる。すべてのビューを、すべての補強に対して不変な共通の埋め込み空間に投影する代わりに、共通の表現は、時間に対して可変または不変な複数の埋め込み部分空間に投影される（図2参照）。したがって、共有表現には時間変化する特徴と不変な特徴の両方が含まれることになり、時間変化を伴うかどうかにかかわらず、リモートセンシングの下流に効率的に転送される。

3.2.1　ビュー・ジェネレーション

参照画像（クエリー）が与えられると、季節的および人工的な補強を施した複数の正対（キー）を生成する。

ション。T を 、 ランダムな切り抜き、カラー・ジッター、色調補正など、よく使われる人為的な補正の集合とする [18]。

ランダム反転。x^t 、$x^{t\,1}$、$x^{t\,2}$ の3つの画像を取得する。これらの画像は、その場所で利用可能なすべての画像の中からランダムに選択される。つまり、$x^q = x^t\,0$ となる。したがって、$x^t\,1$ と $x^t\,2$ は、x^q の季節拡張（または時間ビュー）と考えることができる。

ション、$x^k\,0 = \mathrm{T}\,(x^t\,1\,)$、2番目のキービューには以下のものしか含まれていない。

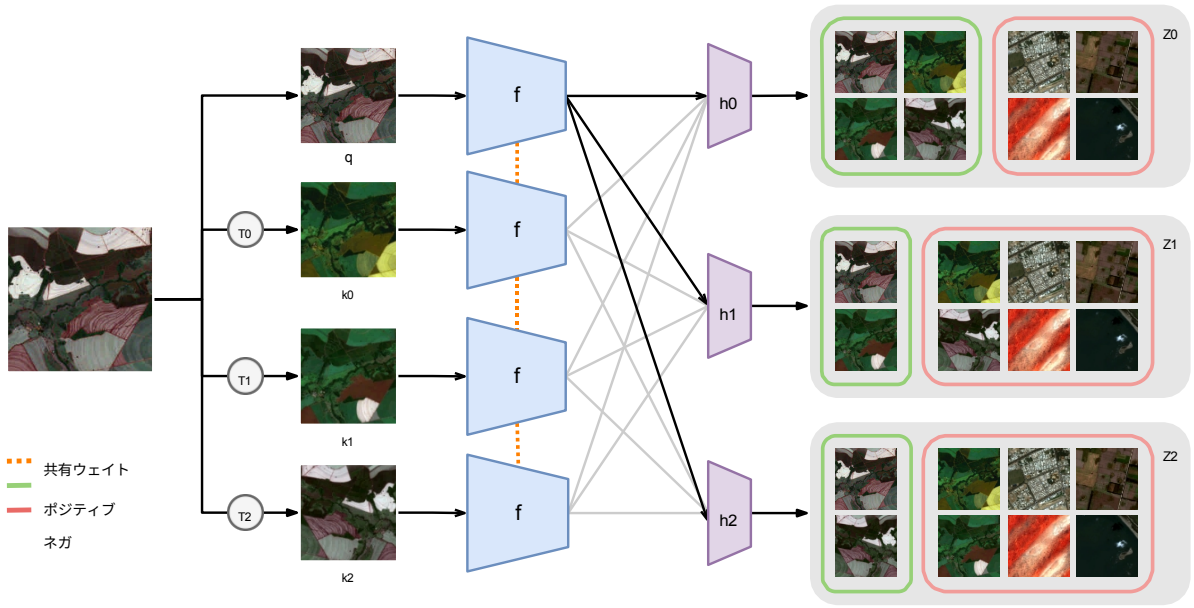

図2.**季節コントラスト法の図。**クエリ画像(q)は時間変換($k0$, $k1$)と合成変換($k0$, $k2$)Tで補強される。エンコーダfによって生成された画像埋め込みは、頭部h_0, h_1, h_2によって3つの異なる部分空間に射影される。緑色のボックスは正のペアを表し、赤色のボックスは負のペア（つまり他の場所の画像を含む）を表す。部分空間Z_0 はすべての変換に対して不変であり、したがってすべてのキーはクエリと同じクラスに属する。Z_1は季節的な補強に対して不変であり、Z_2は合成補強に対して不変である。

季節的な補強、$x^k 1 = x^t 2$ 、第3のビューは人工的な補強のみ、$x^k 2 = T(x^t 0)$ を含む。

3.2.2 複数の埋め込み部分空間

クエリとキービューは、ニューラルネットワークfによって、共通の埋め込み空間$V \in R^d$ 内の表現v^q , $v^k 0$, $v^k 1$, $v^k 2$にエンコードされる。
は、3つの異なる部分空間 Z_0 , Z_1 , $Z_2 \in R^d$ に投影される。
非線形射影頭部 h_0 , h_1 , h_2 , ここで $h_i : V' \to Z_i$. 最近の文献[44]に従って、埋め込み部分空間
はl_2 -正規化され、実質的に単位超球に制限される。

埋め込み部分空間Z_0 は、すべてのオーグメンテーションに対して不変であり、Z_1 は、季節的なオーグメンテーションに対して不変であるが、人工的なオーグメンテーションに対しては可変であり、Z_2 は、人工的なオーグメンテーションに対して不変であるが、季節的なオーグメンテーションに対しては可変である。すなわち、Z_0 、すべての埋め込みz^i 、Z_1 では、z^q と $z^k 1$ だけを一緒に引っ張って押す。

$z^k 0$ と $z^k 2$ は別として、Z_2 では z と $z^k 2$ だけが必要である。

この埋め込み部分空間における他のインスタンスの埋め込み。埋め込み部分空間Z_1 では、クエリ z^q は $z^k 1$ であり、一方、負のペアは他の埋め込みである。
この埋め込み部分空間のインスタンスに加え、$z^k 0$ と $z^k 2$ がある。

$z1$と$z1$は*同じ*インスタンスに由来するが、異なるアーティフィシャル・オーギュメントを持つため、$z1$*に対して*より難しいネガティブ・ペアである。埋め込みにおける正負のペア
ディング空間 Z_2 は Z_1 に類似している。

最終的な学習目標は、すべての埋め込み部分空間の損失の合計である。エンコーダf*は*、一般的な埋め込みにおいて、時間変化する不変な情報を保持しなければならない。
空間 V は，すべての正規化埋め込み部分空間 Z_i の結合された対比的学習目的を最適化するために用いられる．
埋め込み部分空間Z_0 。

転移学習のための表現は、一般的な埋め込み空間Vから取られる。
下流タスクに関する*事前*知識。もし
下流タスクの不変量がわかっている場合、下流タスクの不変量がわかっている場合、下流タスクの不変量がわかっている場合、下流タスクの不変量がわかっている場合、下流タスクの不変量がわかってい

る場合
1 1 2 2
z^k 0とz^k 1から引き寄せられ、押し離される。[2]これを視覚的に表したのが図2である。

コントラスト学習目的は、式1に基づいて各埋め込み部分空間上で最適化される。
正（と負）のペアの定義は、符号化される分散（と分散）に依存する。Z_0 では、正
クエリz^q に対するペアはz^k 0である。
0 0

特定の埋め込み部分空間 Z_i から、表現が抽出される可能性がある。

4. 実験

本研究では、3つの下流タスク、すなわち2つの土地被覆分類タスクについて、学習された表現を評価した、

事前トレーニング	バックボーン	10万枚の画像				1M画像			
		リニアプロービング		微調整		リニア	プロービング		
		10	100%	10%	100%	10		微調整 100	
								10%	100%
ランダムに開始する。	レスネット-18	43.05	45.95	68.11	79.80	43.05	45.95	68.11	79.80
イメージネット		65.69	66.40	78.76	85.90	65.69	66.40	78.76	85.90
モコ-v2	レスネット-18	69.70	70.90	78.76	85.17	69.28	70.79	78.33	85.23
MoCo-v2+TP		70.20	71.08	79.80	85.71	72.58	73.60	80.68	86.59
SeCo（当社製）		**74.67**	**75.52**	**81.49**	**87.04**	**76.05**	**77.00**	**81.86**	**87.27**
ランダムに開始する。	レスネット-50	43.95	46.92	69.49	78.98	43.95	46.92	69.49	78.98
イメージネット		70.46	71.82	80.04	86.74	70.46	71.82	80.04	86.74
モコ-v2	レスネット-50	71.85	73.27	79.23	85.79	73.71	75.65	80.08	86.05
MoCo-v2+TP		72.61	73.91	79.04	85.35	74.50	76.32	80.20	86.11
SeCo（当社製）		**77.49**	**79.13**	**81.72**	**87.12**	**78.56**	**80.35**	**82.62**	**87.81**

表1.BigEarthNet土地被覆分類タスクの平均精度。結果は、異なる事前学習アプローチと異なるResNetバックボーンをカバーしています。また、ラベルなし事前学習セットのサイズを100kから1Mの間で、BigEarthNet学習セットのサイズを10%から100%の間で変化させた場合の影響も調べました。

季節の変化に対して不変でなければならない表現と、季節の変化に対して変化する表現でなければならない変化検出タスク。

プリトレーニングの実装の詳細 我々は、Mo-mentum Contrast (MoCo-v2) [6]を、その最先端の性能とメモリ効率の組み合わせから、本手法のバックボーンとして採用する。我々はMoCo-v2と同じ人工補強、すなわちカラージッタリング、ランダムグレースケール、ガウスぼかし、水平反転、ランダムサイズ変更トリミングを適用する。特徴抽出器としてResNet[17]アーキテクチャを使用し、各埋め込み部分空間に対してReLU活性化と128次元出力を持つ2層MLPヘッドを使用する。また、各埋め込み部分空間に対して個別のキュー[18]を使用し、一度に16,384個の負の埋め込みを含む。バッチサイズ256で、200エポックの事前学習を行う。モメンタム0.9、ウェイト減衰1e-4のSGDオプティマイザを使用。初期学習率を0.03に設定し、エポックの60%と80%でそれを10で割る。温度スケーリングτは0.07である。収集されたデータセットには最大12のスペクトルバンドが含まれるが、より一般的なモダリティであるため、本研究ではRGBチャンネルに注目する。

方法 我々は教師なし学習アプローチを、ランダム初期化、ImageNet教師あり事前学習、自己教師あり事前学習を含むいくつかのベースラインと比較する。後者については、時間情報を利用しない教師なしデータセットに対するMoCo-v2の事前学習の結果を提供する。この場合、データセットの長さは地理的位置の数ではなく画像の総数に依存するため、事前学習エポック数を位置ごとの画像数で割る。また、正ペア（MoCo-v2+TP）を生成するための時間情報を活用したデータセットでのMoCo-v2事前学習の結果、すなわち正ペア（MoCo-v2+TP）を生成するための時間情報を活用したデータセットでのMoCo-v2事前学習の結果も示す。

画像ペアは同じ場所から異なる時間に送られてきたものであり、MoCo-v2人工補強は空間的に整列された画像ペアに適用される（Ayushら[1]と同様）。我々は、線形プロービング（エンコーダをフリーズさせ、分類器のみをトレーニングする）とファインチューニング（エンコーダと分類器の両方のパラメータを更新する）を用いて、全ての手法を評価する。

4.1. BigEarthNetにおける土地被覆の分類

BigEarthNet [38]は、Sentinel-2 [10]画像のチャレンジングで大規模なマルチスペクトルデータセットであり、我々の教師なしデータセットのものと同様のセンサーで撮影され、すなわち12個の周波数チャンネル（RGBを含む）が提供されている。このデータセットは、2017年6月から2018年5月の間にヨーロッパ10カ国で取得された125のSentinel-2タイルから構成され、それぞれ590,326の非重複画像パッチに分割されている。
ピクセルあたりの解像度は10m、20m、60mで、1.2×1.2kmのエリアをカバーしている。の約12％を廃棄した。
季節的な雪、雲、雲の影で完全に覆われたパッチ。これは各画像が複数の土地被覆クラスによって注釈されたマルチラベルデータセットであるため、平均平均精度（mAP）で下流の性能を測定する。我々は[39]で導入された新しいクラス命名法を採用し、[30]で提案されたのと同じ訓練／評価分割を使用する。

実装の詳細 我々は、su-pervised learningで線形分類層を学習することにより、学習された表現を評価する。ResNetのバックボーンを事前に訓練された表現で初期化し、中間表現からクラスロジットにマッピングする1つの完全連結層を追加する。バッチサイズ1024で100エポックのネットワークを微調整し、各実行の最良の検証結果を報告する。デフォルトのハイパーパラメータでAdamオプティマイザを使用。線形プロービングでは、初期学習率を1e-3に設定し、完全なファインチューニングでは、初期学習率を1e-3に設定する。

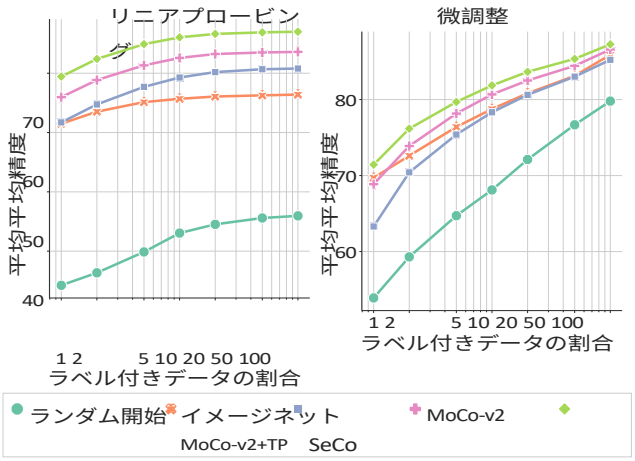

図3.BigEarthNet上でのラベル効率的な土地被覆分類。ResNet-18のバックボーンを使用。

を1e-5に設定した。学習中、エポックの60%と80%で学習率を10で割る。

定量的な結果 表1は、BigEarthNet上でのSeCo事前学習の精度を他の事前学習法と比較したものである。比較は、バックボーン、事前学習イメージの数、BigEarthNetのラベル付きデータのパーセンテージを変えて、線形プロービングまたはファインチューニングで行う。線形プロービングでは、SeCoが一貫してMoCo-v2+TPを上回ることが確認された。また、Temporal pos-itives(TP)はMoCo-v2の性能を僅差で向上させる。さらに、SeCoの特徴量はImageNetで事前に訓練された特徴量よりも有意に向上することがわかり、リモートセンシングと自然画像のドメイン間にはギャップがあるという我々の仮説を確認した。また、BigEarthNet学習セット全体に対してImageNet事前学習済み特徴抽出器を微調整することで、このギャップが減少することもわかった。それにもかかわらず、1Mの画像とResNet-50のバックボーンを用いた場合、SeCoの特徴量はImageNetの特徴量よりも1.1%高い精度を達成した。我々の知る限り、教師なし手法が100%のラベルを持つBigEarthNetにおいて、Ima- geNetの事前学習よりも高い精度を得たのはこれが初めてである。バックボーンサイズに関しては、ResNet-18とResNet-50では、ネットワーク全体を微調整した場合よりも、線形プロービングを行った場合の方が、精度の差が大きいことが確認された。また、バックボーンに関係なく、1Mの画像で事前学習した方が、より高い性能が得られることがわかった。全ての場合において、BigEarthNetのラベル付きデータの10%のみを使用した場合、SeCoはベースラインよりも効率的であることが分かる。

サンプリング	リニアプロービング10	100%	微調整10	100%
ガウシアン	**74.67**	**75.52**	**81.49**	**87.04**
ユニフォーム	71.63	72.59	79.65	85.75

表 2.SeCoデータセットのサンプリング戦略の比較。ResNet-18のバックボーンを使用。

教師あり学習はラベル付きデータの量とともに増加するが、自己教師あり手法の差は大きく変わらない。すべてのラベル付きデータの割合について、SeCoは、MoCo-v2+TPと比較して、、一定の約4%の改善ギャップを達成している。微調整を行うと、性能は

図 3 は、BigEarthNet の**ラベル**付きデータのパーセンテージを変えた場合の、SeCo と各種ベースラインの線形プロービングと微調整のパフォーマンスを示している。線形プロービングでは、BigEarthNetのラベルのわずか1%で、SeCoはラベルの100%でImageNetの事前学習を上回り、ラベルの20%でMoCo-v2と一致することが分かった。また、ImageNetの事前学習と自己学習との間のギャップは、ImageNetの事前学習と自己学習との間のギャップよりも小さいことがわかった。

ラベル付きデータの割合を増やすと、自己教師付き手法とImageNetの差は縮まる。とはいえ、SeCoはすべてのベースラインよりもラベル効率が高く、利用可能なすべてのラベルを使用したImageNetの事前学習と、わずか50%のラベルで性能が一致する。

位置情報のサンプリング戦略に関するアブレーション リモートセンシング表現の事前学習用に、未修正画像を収集するためのサンプリング戦略の有効性を評価するために、地球の位置情報が大陸内から単一形式でサンプリングされたSeCoデータセットの代替バージョンをダウンロードした。このアプローチに従って10万枚の画像をダウンロードし、ResNet-18をSeCo手法で事前学習する。表2はBigEarthNetダウンストリームタスクにおいて、各サンプリング方式を用いた場合の伝達学習のパフォーマンスを比較したものである。人間の居住地を中心としたガウス分布（3.1節参照）の混合からサンプリングされた画像で事前学習されたSeCo表現が、画像を一様にサンプリングするよりも、下流での性能が良いことがわかる。これは、人里離れた地域は人間の活動により多様である傾向があるため、収集された画像には良い表現を学習するための情報がより多く含まれているためであると主張する。

4.2. EuroSATによる土地被覆分類

EuroSAT [19]もまた、Sentinel-2衛星画像を用いた土地利用と土地被覆の分類という課題に取り組んでいる。画像はヨーロッパ34カ国に対応し、異なる土地利用に対応する10のクラスで構成されている。各クラスは2,000〜3,000枚の画像で構成され、合計27,000枚のラベル付き画像がある。画像
サイズは64 × 64ピクセルで、640 × 640 mの領域をカバーしている。我々は
30]で提案されたものと同じtrain/val分割である。

実装の詳細 このタスクでは、教師あり学習で線形分類器を学習することで、学習された表現も評価する。ResNet-18のバックボーンを事前に学習した表現で初期化し、その上に1つの完全連結層を追加する。この場合、 1Mの衛星画像で事前に訓練された表現でバックボーンを初期化する（ランダムな重みを使用する場合、またはロードする場合を除く）。

事前トレーニング	精度
ランダムに開始する。	63.21
イマジネット	86.44
モコ-v2	83.72
モコv2+TP	89.51
SeCo（我々）	**93.14**

表3.EuroSAT土地被覆分類タスクの微調整精度。ResNet-18のバックボーンを使用。

ImageNetで事前に訓練されたモデル)。バックボーンの重みを凍結し、バッチサイズ32で100エポックの間分類器を訓練し、各実行の最良の検証精度を報告する。デフォルトのハイパーパラメータでAdamオプティマイザを使用し、初期学習率を1e-3に設定し、エポックの60%と80%で10で割る。

定量的な結果 表 3 は、SeCo 表現の線形プローブ精度を異なるベースラインと比較したものである。SeCo は ImageNet の事前学習よりも 6.7%、MoCo-v2+TPよりも3.6%高い精度を達成していることがわかる。これらの結果は、学習された表現がBigEarthNetで有効であるだけでなく、EuroSATのような他のリモートセンシングデータセットでも有効であることを示している。

4.3. 衛星の変化検出

Onera Satellite Change Detection（OSCD）データセット[8]は、Sentinel-2からの24組のマルチスペクトル画像で構成されている。画像は2015年から2018年にかけて、様々なレベルの都市化が見られる世界中の場所から記録されたもので、都市の変化が確認できる。各ロケーションには、13のSentinel-2スペクトルバンドすべてをカバーする整列されたペアが含まれている。画像は10m、20m、60mの間で空間分解能が異なり、およそ
600×600ピクセル、解像度10m。目的は、異なる日付の衛星画像間の変化を検出することである。

事前トレーニング	精度	リコール	F1
ランダムに開始する。	**70.53**	19.17	29.44
イマジネット	**70.42**	25.12	36.20
モコ-v2	64.49	30.94	40.71
モコv2+TP	69.14	29.66	40.12
SeCo（我々）	65.47	**38.06**	**46.94**

すべての訓練画像と検証画像のペアに対して、ピクセルレベルの変化グランドトゥルースを提供する。Daudtら[8]が提案したのと同じ訓練/検証の分割を使用する：トレーニング用画像は14枚、検証用画像は10枚である。画像セグメンテーションの文献で一般的なように、F1スコアで下流の性能を測定する。

実装の詳細 2つの異なるタイムスタンプで与えられた場所からの画像のペアごとに、Daudtら[7]と同様の手順でセグメンテーションマスクを生成する。まず、ResNet-18バックボーンが各画像から特徴を抽出する。バックボーンネットワークでダウンサンプリングが行われるたびに特徴量を保持する。そして、各ペアの2つの特徴量の差の絶対値を計算し、その特徴量の差をU-Net[34]の入力として使用し、バイナリー・セグメンテーション・マスクを生成する。バックボーン・ネットワークは、1Mの衛星画像で事前に訓練された表現で初期化される。には

表 4.Onera Satelliteの変化検出タスクの微調整結果。ResNet-18を使用。

オーバーフィッティングを避けるため、バックボーンを凍結し、U-Netの重みのみを訓練し、U-Netの各アップサンプリング層の後に0.3のドロップアウト率を追加し、ランダムな水平反転と90° の回転で訓練画像を補強する。さらに、OSCDデータセットの画像はサイズが可変であるため、以下のような非重複パッチに分割する。
96×96ピクセル。バッチサイズ32で100エポックのデコーダを訓練し、以下の検証セットの結果を報告する。
change "クラスの視点である。重みの減衰を1e-4とするアダム・オプティマイザを用いる。初期学習率を1e-3に設定し、各エポックで0.95の乗法係数で指数関数的に減少させる。

定量的結果 表4はSeCoをran-dom初期化、ImageNet事前学習、MoCO-v2、MoCo-v2+TPと比較したものである。SeCo初期化はすべてのベースラインよりも高いリコールとF1スコアを達成していることがわかる。特に、SeCoはMoCo-v2+TPよりもF1スコアが6.8%高い。これはMoCo-v2+TP表現が時間的変化に対して不変であるためであろう。興味深いことに、SeCoもMoCo-v2も、異なるタイムスタンプの同じ位置からの画像パッチを負のペアとみなす（つまり、学習された表現が時間に対して変化する）が、SeCoの方が6.2%高いF1スコアを保持している。このことは、複数の埋め込み部分空間によって、SeCoが時間的変化から画像の増大を分離することで、時間的変化をより効果的に検出できることを示している。

定性的結果 図4は、OSCD検証セットから2つのサンプルについて、我々の手法とすべてのベースラインによって生成された変化検出マスクを比較したものである。SeCoの事前学習は、過剰な偽陰性を発生させることなく、変更されたピクセルの多くをカバーする、より高品質なマスクを生成することがわかった。また、MoCo-v2の性能には、温度情報(TP)を利用した場合と利用しなかった場合で、いくつかの相違があることがわかった。これは、各アプローチで時間的不変性の扱いが異なるため、画像の違いがより人為的な補強や時間的変化に似ているためであると考えられる。SeCoは、時間変化する不変要素を保持する表現を学習することで、この問題を克服する。

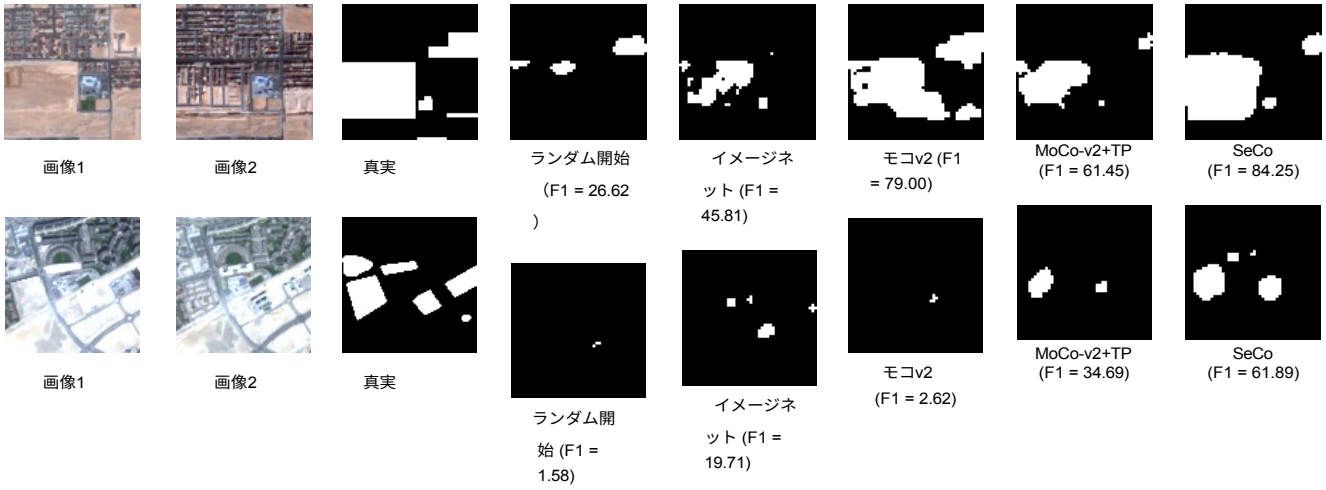

図4.Onera Satelliteの変化検出タスクの定性的結果の比較。各行には、検証サンプルに対する入力画像、グランドトゥルースマスク、生成された変化検出マスクが含まれる。

5. 関連作品

非教師ありデータからの学習 最近の非教師あり特徴学習では、ImageNetのような小規模データセットか、高度にキュレーションされたデータセットのどちらかに焦点が当てられているが、一方で、キュレーションされていない生のデータセットを使用すると、転送タスクで評価したときに特徴の質が低下することが分かっている[9, 2]。Caronら[3]は、キュレーションされていないデータで訓練された教師なしメソッドの性能を向上させるために、クラスタリングを活用した自己教師ありアプローチを提案している。他の手法では、ハッシュタグ[23, 40, 27]、ジオロケーション[45]、ビデオ構造[14]などのメタデータをノイズの多い監視ソースとして利用する。我々の研究では、リモートセンシングデータの地理的・時間的情報を活用し、未修正データセットから教師なし表現を学習する。

最近の自己教師付き対比学習法は、異なる画像拡張に対して不変であることを学習することで、印象的な伝達可能な視覚表現を生成することができる。しかし、これらの方法は、暗黙のうちに特定の表現的不変性を仮定しており、下流のタスクがこの仮定に違反すると、うまく機能しないことがある。Xiaoら[47]はLeave-one-out Contrastive Learning (LooC)を提案している。これは、1つの補強を除く全ての補強に対して不変な埋め込み空間を別々に構築することで、変化する不変な要素を捉えることができる視覚表現を生成する、多補強対照学習フレームワークである。我々の研究では、同様のアプローチを用いて、リモートセンシング画像に存在する季節的変化に対して変化し不変な表現を学習する。

リモートセンシングにおける教師なし学習 自然画像データセット（例えばImageNet）では教師なし学習が広く研究されているが、リモートセンシング領域ではこのサブフィールドはまだ未解明である。これは、地球観測におけるリモートセンシングの重要性、容易に利用可能な膨大なデータ量、そして、リモートセンシングの特徴である

衛星画像の特性。例えば、Jean et al. [22]は、正負のペアをサンプリングするために画像の地理情報を使用し、三重項損失に基づく前文タスクを構築する。Uzkentら[42]は、地理参照されたウィキペディアの記事と、対応する場所の衛星画像をペアリングし、画像から記事の特性を予測することで表現を学習する。Vincenziら[43]は、リモートセンシング画像の多スペクトル性を利用して、他のスペクトルバンドから可視色を再構成するカラー化タスクを構築している。また、Ayushら[1]は、衛星画像の時間情報を利用して、正対を生成し、対比目的を学習することを提案している。しかし、彼らの表現は時間的変化に対して常に不変であるため、時間的変化を伴う下流のタスクでは不利になる可能性がある。我々は、時変不変な情報を保持した表現であるマルチオーグメンテーションコントラスト学習を用いることで、この問題を克服する。

6. 結論

我々は、リモートセンシング画像のための新しい転移学習パイプラインであるSeasonal Contrast（SeCo）を発表した。SeCoは、データ収集戦略と、このデータを活用する自己教師付き学習アルゴリズムから構成される。まず、複数のタイムスタンプで人口密集地域周辺をサンプリングし、多様な衛星画像セットを提供する。次に、季節変化を考慮し、豊富で転送可能なリモートセンシング表現を学習するために、マルチ拡張コントラスト学習法を拡張する。

SeCoと一般的なImageNetの事前学習およびMoCoの事前学習を、異なるバックボーンとデータセットサイズを用いて収集したデータで比較した。その結果、SeCoはBigEarthNet、Eu-roSAT、OSCDの各タスクにおいて、考慮したベースラインを上回る結果を得た。したがって、ImageNetのような標準的なデータセットやMoCoのようなアルゴリズムを用いた事前学習よりも、ドメインに特化した教師なし事前学習の方が、再モータセンシングアプリケーションには効果的であると結論付けた。

参考文献

[1] K.アユシュ、B.ウズケント、C.メン、M.バーク、D.ロベル、そして

S.S. Ermon.地理を意識した自己教師付き学習. *arXiv preprint arXiv:2011.09980*, 2020.5, 8

[2] M.Caron, P. Bojanowski, A. Joulin, and M. Douze.Deep clustering for unsupervised learning of visual features.*Proceedings of the European Conference on Computer Vi- sion (ECCV)*, pages 132-149, 2018.8

[3] M.Caron, P. Bojanowski, J. Mairal, and A. Joulin.非キュレーションデータにおける画像特徴の非スーパービジョン事前学習。*Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2959-2968, 2019.8

[4] M.このような場合、「視覚的な特徴」の学習は、「視覚的な特徴」の学習と、「視覚的な特徴」の学習と、「視覚的な特徴」の学習とに分けられる。*arXiv:2006.09882*、2020.2

[5] T.(1)視覚的表現の対比学習のためのシンプルなフレームワーク。A sim-ple framework for contrastive learning of visual representsen- tations.In *International conference on machine learning*, pages 1597-1607.PMLR, 2020.2

[6] X.(注1)本論文は、本論文の一部である。運動量対比学習によるベースラインの改善. *arXiv preprint arXiv:2003.04297*, 2020.5

[7] R.C. Daudt, B. Le Saux, and A. Boulch.変化検出のための完全畳み込みシャムネットワーク。In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 4063-4067.IEEE, 2018.7

[8] R.C. Daudt, B. Le Saux, A. Boulch, and Y. Gousseau.Urban change detection for multispectral earth observation using convolutional neural networks.In *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Sympo- sium*, pages 2115-2118.IEEE, 2018.2, 7

[9] C.Doersch、A. Gupta、A. A. Efros.Unsupervised vi- sual representation learning by context prediction.In *Proceedings of the IEEE international conference on computer vision*, pages 1422-1430, 2015.2, 8

[10] M.ドルシュ、U.デル・ベロ、S.カルリエ、O.コリン、V.フェルナンデス、

F.Sentinel-2：ESAのgmes運用サービスのための光学高解像度ミッション。*環境のリモートセンシング*, 120:25-36, 2012.1, 2, 3, 5

[11] F.フィリッポニセンチネル2時系列を利用した国レベルでの焼失地域の地図作成：2017年イタリア山火事のケーススタディ。*Remote Sensing*, 11(6):622, 2019.1

[12] G. M. フーディ。熱帯林環境のリモートセンシング：持続可能な開発のための環境資源のモニタリングに向けて。*International journal of remote sensing*, 24(20):4035-4046, 2003.1

[13] S.Gidaris, P. Singh, and N. Komodakis.*arXiv preprint arXiv:1803.07728*, 2018.2

[14] D.ゴードン、K.エハニ、D.フォックス、A.ファルハディ。世界が通り過ぎるのを見る： *arXiv preprint arXiv:2003.07990*, 2020.8

[15] N.ゴレリック、M.ハンチャー、M.ディクソン、S.イリューシチェンコ、
D.Thau, and R. Moore.グーグルアースエンジン：みんなのための惑星規模の地理空間分析。*Remote sensing of Environment*, 202:18-27, 2017.3

[16] J.-B.Grill, F. Strub, F. Altche´, C. Tallec, P. H. Richemond、
E.ブカツカヤ、C. ドアシュ、B. A. ピレス、Z. D. グオ、M. G. アザール、他 潜在を自分でブートストラップする： *arXiv preprint arXiv:2006.07733*, 2020.2

[17] K.He, X. Zhang, S. Ren, and J. Sun.Deep residual learning- ing for image recognition.この論文では、*画像認識のための深い残差学習について述べている*。5

[18] K.He,H.Fan,Y.Wu,S.Xie,R.Girshick。このような視覚表現の学習は、教師なし視覚表現学習のためのモーメンタムコントラストである。このような視覚表現の学習は、学習された視覚表現と学習された*視覚表現との間に、どのような関係が存在するのかを明らかにする*。2, 3,
5

[19] P.ヘルバー、B.ビシュケ、A.デンゲル、D.ボース。
ユーロサット： A novel dataset and deep learning benchmark for land use and land cover classification.*IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2019.1, 2, 6

[20] C.イポリーティ、L.カンデロロ、M.ギルバート、M.ゴフレード、
G.マンチーニ、G.クルチ、S.ファラスカ、S.トラ、A.ディ・ロレンツォ
M.多変量クラスタリングアプローチに基づくイタリアにおける生態学的地域の定義：多変量クラスターアプローチに基づくイタリアにおける生態学的地域の定義：標的化された媒介感染症サーベイランスに向けた第一歩。*PloS one*, 14(7): e0219072, 2019.1

[21] N.ジーン、M.バーク、M.謝、W.M.デイビス、D.B.ロベル、そして
S.エルモン衛星画像と機械学習を組み合わせた貧困予測。*Science*, 353(6301):790-794, 2016.3

[22] N.(注1)本データはこの書籍が刊行された当時に掲載されていたものです。Tile2vec：空間的に分散したデータのための教師なし表現学習。*Proceedings of the AAAI Con- ference on Artificial Intelligence*, volume 33, pages 3967- 3974, 2019.8

[23] A.Joulin, L. Van Der Maaten, A. Jabri, and N. Vasilache.弱教師付き大規模データからの視覚的特徴の学習。*European Conference on Computer Vision*, pages 67-84. Springer, 2016.8

[24] I. ララジ、P .ロドリゲス、O .マナス、K .レンシンク、M.ロー、L.カーズマン、W.パーカー、D.バスケス、そして

D.Nowrouzezahrai.CT画像におけるcovid-19セグメンテーションのための弱教師付き一貫性ベース学習法.2021.1

[25] I.H.ララジ、D.バスケス、M.シュミット。マスクはどこにあるのか：画像レベルの監視によるインスタンスセグメンテーション。2019.1

[26] I.H.ララジ、R.パルディナス、P.ロドリゲス、D.バスケス。Looc：重なり合う物体をカウント監視でローカライズする。2020.1

[27] D.マハジャン、R.ギルシック、V.ラマナタン、K.ヘー、M.パルリ、
Y.Li、A. Bharambe、L. Van Der Maaten。弱教師付き事前トレーニングの限界を探る。*Proceedings of the European Conference on Computer Vision (ECCV)*, pages 181-196, 2018.8

[28] I.I. Misra and L. v. d. Maaten. 前文不変表現の自己教師付き学習。*IEEE/CVF Conference of Computer Vision and Pattern Recognition*, pages 6707-6717, 2020.2

[29] D.J. ムラ精密農業におけるリモートセンシングの25年：主な進歩と残された知識のギャップ。
バイオシステムズ・エンジニアリング, 114(4):358-371, 2013.1

[30] M.Neumann, A. S. Pinto, X. Zhai, and N. Houlsby.*arXiv preprint arXiv:1911.06721*, 2019.1, 5, 6

[31] M.Noroozi and P. Favaro.ジグソーパズルを解くことによる視覚表現の教師なし学習。In *European conference on computer vision*, pages 69-84.Springer, 2016.2

[32] A. v. d. Oord, Y. Li, and O. Vinyals.*ArXiv preprint arXiv:1807.03748*, 2018.2, 3, 4

[33] D.ロルニック、P.L.ドンティ、L.H.カーク、K.コチャンスキー、
A.Lacoste, K. Sankaran, A. S. Ross, N. Milojevic- Dupont, N. Jaques, A. Waldman-Brown, et al. Tack- ling climate change with machine learning.*arXiv preprint arXiv:1906.05433*, 2019.1

[34] O.U-net.生体画像セグメンテーションのための対話型ネットワーク：U-net: Convo-lutional networks for biomedical image segmentation.In *International Conference on Medical image computing and computer-assisted intervention*, pages 234-241.Springer, 2015.7

[35] D.P.ロイ、M.A.ウルダー、T.R.ラブランド、C.E.ウッドコック、
R.Landsat-8: Science and Product Vision for Terrestrial Global Change Research.*Remote sensing of Environment*, 145:154-172, 2014.2

[36] O.O.ルサコフスキー、J.デン、H.スー、J.クラウス、S.サティーシュ、
S.Imagenet large scale visual recognition challenge (2014). *arXiv preprint arXiv:1409.0575*, 2014.2, 3

[37] G. J. Schumann、G. R. Brakenridge、A. J. Kettner、R. Kashif、E. Niebuhr。地球観測データとプロダクトによる洪水災害対応の支援：リモートセンシング, 10(8).*Remote Sensing*, 10(8):1230, 2018.1

[38] G. Sumbul, M. Charfuelan, B. Demir, and V. Markl.Bigearthnet：リモートセンシング画像理解のための大規模ベンチマークアーカイブ。In *IGARSS 2019-2019 IEEE International Geoscience and Remote Sensing Symposium*, pages 5901-5904.IEEE, 2019.1, 2, 5

[39] G・スンブル、J・カン、T・クロイツィガー、F・マルセリーノ、H・コスタ

P.Benevides、M. Caetano、B. Demir。リモートセンシング画像理解のための新しいクラス命名法を用いたBigearthnetデータセット。5

[40] C.Sun, A. Shrivastava, S. Singh, and A. Gupta.Revisiting unreasonable effectiveness of data in deep learning era.*Proceedings of the IEEE international conference on com-puter vision*, pages 843-852, 2017.8

[41] Y.Tian, D. Krishnan, and P. Isola.Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019.2

[42] B.ウズケント、E.シーハン、C.メン、Z.タン、M.バーク、

D.Lobell, and S. Ermon.ウィキペディアを用いた地球規模の衛星画像の解釈の学習。 *arXiv preprint arXiv:1905.02506*, 2019.1, 8

[43] S.ヴィンチェンツィ、A.ポレッロ、P.ブゼガ、M.チプリアーノ、P.フロンテ、

R.R. Cuccu, C. Ippoliti, A. Conte, and S. Calderara.地球観測画像のための自己教師付き表現の学習 . *arXiv preprint arXiv:2006.12119*, 2020.8

[44] T.Wang and P. Isola.(1)超球面上の整列と一様性を通して、対照的な表現学習を理解する。In *International Conference on Machine Learning*, pages 9929-9939.PMLR, 2020.4

[45] T.Weyand, I. Kostrikov, and J. Philbin.畳み込みニューラルネットワークによる惑星写真ジオロケーション。In *European Conference on Computer Vision*, pages 37-55.Springer, 2016.8

[46] Z.Wu, Y. Xiong, S. X. Yu, and D. Lin.Unsupervised feature learning via non-parametric instance discrimination.In *Pro- ceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3733-3742, 2018.2

[47] T.X. Wang, A. A. Efros, and T. Darrell.対照学習において対照的であってはならないもの。2, 3, 8

[48] R.Zhang, P. Isola, and A. A. Efros.Colorful image colorization.In *European conference on computer vision*, pages 649-666.Springer, 2016.2