教師なし学習のための自己教師学習型学習 正規化カットを用いた物体検出

CVPR22論文の著者バージョン

Yangtao Wang1 、Xi Shen2,3, 、シェル徐胡4 、元元5 、ジェームズ•L•クロウリー1 、ドミニク•ヴォーフリーダズ1
[1] グルノーブル•アルプ大学、CNRS、グルノーブルINP、LIG、38000 グルノーブル、フランス
[2] テンセントAIラボ [3] LIGM (UMR 8049) - エコール•デ•ポン、UPE
[4] サムスンAIセンター、ケンブリッジ [5] MIT CSAIL

抽象的な

自己蒸留損失（DINO)を用いた自己教師学習によって学習された変換装置は、目立つ前景オブジェクトを強調表示する注目マップを生成することが示されています。
この論文では、グラフベースの手法を用いて
オブジェクトを発見するための自己教師ありトランスフォーマーの特徴
画像から視覚トークンはノードとして表示されます
接続性を表すエッジを持つ重み付きグラフ
トークンの類似度に基づいてスコアリングを行う。そして、前景オブジェクトは、正規化されたグラフカットを用いて自己相似領域をグループ化することで分割できる。グラフカットを解く。
一般化固有分解を用いたスペクトルクラスタリングを用いた問題を解いて、2番目に小さい固有ベクトルがその絶対値から切断解を与えることを示す。
トークンが前景オブジェクトに属する可能性を示します。

このアプローチはシンプルであるにもかかわらず、
教師なしオブジェクト発見のパフォーマンス :我々は
最新のLOSTと比較して、VOC07でそれぞれ6.9%、8.1%、8.1%の改善が見られました。
VOC12、COCO20Kなど。性能はさらに向上する可能性がある。
第二段階のクラス非依存検出器（CAD)を追加することで改善されました。提案手法は簡単に拡張できます。
教師なしの顕著性検出と弱教師ありの
物体検出。教師なしの顕著性検出では、
ECSSD、DUTS、
DUT-OMRON をそれぞれ最先端技術と比較します。
弱教師付き物体検出では、CUBとImageNetで競争力のあるパフォーマンスを達成しました。私たちのコードは
詳細はこちら： https://www.m-psi.fr/Papers/TokenCut2022/

1 はじめに

物体検出は、ロボット工学、自動運転、交通監視、製造、そして、

具現化された人工知能[21, 63, 64]。しかし、

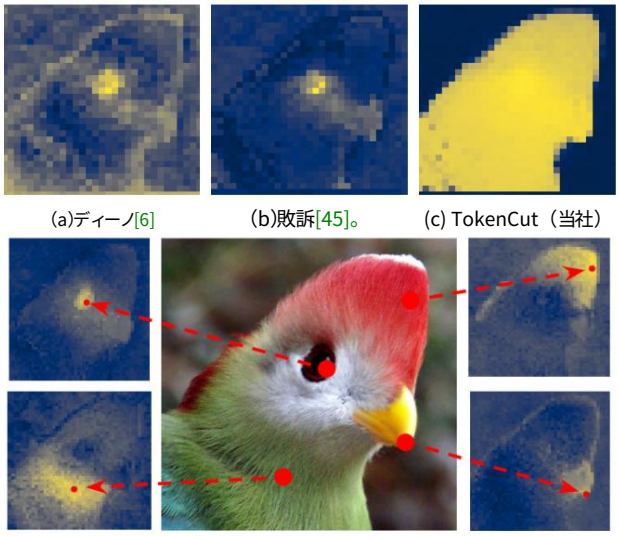


図1: クラストークンの注目マップ
DINO [6] (a) と逆次数のマップは、
LOST [45] (b) は前景と背景の分離にノイズが多い。我々の提案する手法は、前景と背景をきれいに分離する。
注目オブジェクトをセグメント化するために使用できるマップ（c）。
異なるパッチに関連付けられた注目マップを考慮すると、
物体のさまざまな領域を強調表示する場合 (d)、複数のパッチから注目マップを計算するグラフを構築するのが合理的です。

現在の最先端の物体検出器の性能は、十分な情報を注釈付けするための高コストによって制限されている。
教師あり学習のための訓練データ[33] 。この制限は
転移学習を使用するとさらに明らかになる
事前学習済みの物体検出器を新しいアプリケーションに適応させる
能動学習[1]、半教師あり学習[34]、弱教師あり学習[39]などのアプローチは、より効率的な学習を提供することでこの障壁を克服しようと試みてきたが、成功は限られている。

*責任著者

セス。

本研究では、人間による注釈のない自然画像における物体発見に焦点を当てています。これは重要な
これは多くの下流アプリケーションにとって重要なステップであり、問題となる[56]。オブジェクト検出が不十分だと、全体的なパフォーマンスが低下する可能性がある。
システムパフォーマンス。この問題に対する現在のアプローチでは、バウンディングボックス提案メカニズム[11, 55, 56, 61]を採用し、オブジェクト発見を次のように定式化している。
最適化問題。しかし、このアプローチは
計算コストが高いため[55] 、境界の各ペアは
異なる画像間でのボックス提案を比較する必要があり、最適化がより大きなサイズにスケールできない可能性がある。
データセットは二次計算のオーバーヘッドにより分割されやすい[57]。

トランスフォーマーは最近、
視覚認識のための畳み込みニューラルネットワーク。ViT [18]などの
Vi-sion Transformerは画像パッチを受け入れ、
入力トークンとして、エンコーダーのスタックレイヤーを使用して
トークンを画像レベルのクラスラベルにマッピングするための自己注意。
DINO [6]の最近の結果では、
自己蒸留損失[25]で訓練された注目マップ
最後の層のクラストークンに関連付けられている
目立つ前景領域。しかし、
図1aのような注意マップはノイズが多く、明確ではありません
教師なしオブジェクト検出に使用できること。

LOST [45]では、著者らは
グラフを描き、ノードの逆次数を用いてオブジェクトを分割する。ヒューリスティックなシード拡張戦略を用いてノイズを克服し（図1b） 、単一の境界ボックスを検出する。
前景の物体。それに関連する注目マップは
異なるノードには意味のある情報が含まれていることが多いため、
図1dに示されている。我々はそれが
グラフを低次元部分空間に投影することで、グラフ全体の情報を利用することができる。
固有値分解。我々は、このような投影を正規化カット[43] （Ncut)と組み合わせて使用できることを発見した。
前景/背景のセグメンテーションが大幅に改善されました（図1c）。

本稿では、シンプルだが効果的なTokenCutを提案する。
教師なしオブジェクト検出のためのグラフベースのアプローチ。
自己教師あり学習済みのビジョントランスフォーマーをベースに構築
DINO [6]をバックボーン特徴エンコーダとして利用し、
結果の特徴を持つオブジェクト。
クラストークンでは、すべてのトークンの特徴を使用します。
最後のトークン特徴に基づく無向グラフ
自己注意層では視覚トークンが閲覧される
グラフノードとして、エッジは接続を表す
特徴の類似性に基づいてスコアを算出します。次に、
自己相似領域をグループ化するための正規化されたグラフカットと
前景のオブジェクトを区切る。グラフカットを解く。
一般化固有分解を用いたスペクトルクラスタリングを用いて問題を解いて、2番目に小さい固有ベクトルが尤度を示す切断解を与えることを示す。
トークンが前景オブジェクトに属しているかどうか。私たちのアプローチは
実行時の適応として考えられ、それは
モデルがそれぞれの特定のテスト画像に適応できること
共有トレーニング モデルにもかかわらず。

シンプルであるにもかかわらず、私たちのアプローチは教師なし物体発見を大幅に改善します。この方法は
VOC07では68.8%、72.1%、58.8%を達成した[19]。
VOC12 [20]、 COCO20K [33]をそれぞれ6.9%、8.1%、8.1%上回った。
それぞれ。第2段階のCADを備えたTokenCutはさらに
パフォーマンスは71.4%、75.3%、62.6%に向上します。
それぞれ VOC07、VOC12、COCO20k で、LOST + CAD をそれぞれ
5.7%、4.9%、5.1% 上回ります。

さらに、TokenCutは弱教師ありの物体検出や教師なしの顕著性検出にも容易に拡張できることを示します。弱教師ありの物体検出では、画像レベルの情報のみを用いて物体を検出することが目標です。

注釈。エンコーダをフリーズし、線形微調整を行う。
弱教師画像ラベルを用いた分類器。次に
微調整されたエンコーダから抽出された特徴にTokenCutを適用すると、明らかに改善された結果が得られます。
CUBデータセット[59]とImageNet-1K[14]での競争力のあるパフォーマンスの結果。教師なしサリエンシー
検出には、
提案されたアプローチとバイラテラルソルバー[5]を適用する
前景のエッジを洗練させるための後処理ステップ
領域。結果的に、我々のアプローチはECSSD [44]におけるこれまでの最先端手法を大幅に改善した。
DUTS [60]およびDUT-OMRON [67]。

要約すると、私たちの主な貢献は次のとおりです。

- 我々は、発見するためのシンプルで効果的な方法を提案する。
  自己教師あり学習に基づく画像内の物体認識。この手法は、従来の最先端手法を大幅に上回る性能を示す。

  複数のデータセットでテストした場合の教師なしオブジェクト検出用。
- 提案手法を弱教師あり学習に拡張する
  物体検出と、このシンプルなアプローチで
  競争力のあるパフォーマンスを達成する;
- また、この手法が教師なしの顕著性検出にも使用できることを示しています。結果は、
  TokenCut は、複数のデータセットにおけるこれまでの最先端のパフォーマンスを大幅に向上させます。

2 関連研究

自己教師あり視覚変換装置。ViT [18]は
トランスフォーマーアーキテクチャ[53]は、
画像の効果的なエンコーダーであり、教師あり視覚タスクに有用な特徴を提供する。MoCo-v3 [8]は、ViTが自己教師あり表現学習を提供し、対照的な学習を用いて優れた結果を達成できることを実証した。

最近、DINO [6] は自己蒸留損失[25]を用いて変換子を訓練することを提案し、 ViT が意味的学習に使用できる明示的な情報を含んでいることを示した。
画像のセグメンテーション。BERT [16]、 [32]に触発された。
提案されたMSTは、一部のトークンを動的にマスクする
グローバル画像デコーダーを用いて失われたトークンを復元する
方法を学習する。BERT [16]に倣い、 BEIT [4]は初めて
元の画像を視覚的なトークンにトークン化し、いくつかのトークンをランダムにマスクして、次のものを使用してそれらを復元することを学習します。
変圧器。最近、MAE [23]は高い割合で-

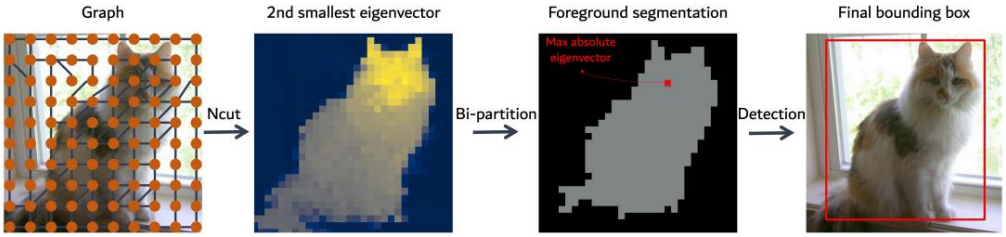


図2: TokenCutアプローチの概要。ノードがトークン、エッジがトランスフォーマーの特徴量を用いたトークン間の類似性です。前景と背景のセグメンテーションはNcut [43]によって解決されました。2番目に小さい固有ベクトルを二分割することで前景の物体を検出できます。

画像の欠落部分を再構成し、非対称エンコーダ/デコーダ。

教師なしオブジェクト発見。画像群が与えられた場合、教師なしオブジェクト発見は、画像群から、複数の画像に表示される類似のオブジェクトを区切ります。いくつかの方法[9, 26, 28, 29, 54]は、セグメント化するように設計されている。画像コレクション内の共通の繰り返しオブジェクトですが、出現頻度に関する強い仮定に基づいて物体の。他のアプローチ[11, 50, 55, 56]では、境界ボックスの提案を使用し、物体の発見を次のように定式化する。最適化問題である。 [57]は、ランキング問題として教師なしオブジェクト発見の新しい定式化を提案し、発見がデータセットにスケールできることを示した。1万枚以上の画像を用いた。最近、LOST [45]は、教師なし学習において最先端の学習法を大幅に改善した。物体検出。LOSTはDINO [6]に基づく自己教師あり変換を用いて特徴を抽出し、単一のオブジェクトを取得するためのヒューリスティックシード拡張戦略地域。私たちの研究はLOST [45]と密接に関連しており、自己教師あり学習も採用している。しかし、特定の注目マップに頼るのではなくノードに対して、我々はグラフベースの手法を提案する。すべてのノードの注目スコアと、Ncut [43]は、より正確なセグメンテーションを得るために、画像オブジェクト。

弱教師あり物体検出。弱教師あり物体検出[22, 68, 69, 72]は、

画像レベルの注釈のみを使用して画像オブジェクトを見つけます。初期のアプローチ[7, 41, 77]は主に[77]で導入されたクラス活性化マップ（CAM)に依存しており、クラス固有の局在マップを生成し、識別領域を見つけていました。いくつかの手法[12, 13, 36, 46, 69, 75]

CAMを改善するために、判別領域を限定し、ネットワークに追加の物体領域を捕捉させる。Cutout [17]やCutMix [70]などのデータ拡張技術は、

分類と
ローカリゼーションのパフォーマンス。いくつかの方法では両方を達成しています分類と位置特定を2つの別々のネットワークを用いて行った[22, 35, 71]。 [71]は位置特定ネットワークを訓練した。[61]によって生成された疑似境界ボックスを使用する。 [71]まず分類器を学習し、その重みを固定して別の分類器を訓練する検出器。 [22]は回帰器と分類器を学習し、

CAMの一貫性を2つの変換間で維持する。これらのアプローチは、

弱教師ありの物体検出に対し、教師なしの物体検出と

トランスフォーマーに基づく弱教師付き物体検出。

教師なしの顕著性検出。教師なし顕著性検出は、画像内の顕著な物体をセグメント化することを目指します。この問題に関するこれまでの研究は[27, 31, 66, 78]は、次のような古典的な手法を使用した。色のコントラスト[10]、特定の背景の事前確率[62]、またはスーパーピクセル[31, 67]。最近では、教師なし深層学習モデル[37, 60, 73]はヒューリスティックな顕著性を組み込んでいる深層CNNを訓練するための疑似グラウンドトゥルースとしての方法モデル。しかし、これらの方法はCNNモデルに依存している。監督下での事前訓練を受けた。 [58]は教師なし大規模GANは訓練中のラベルの使用。以下では、簡単な後処理ステップを組み込むことで私たちの教師なしオブジェクト発見は強力な教師なし顕著性検出のベースライン手法。

3 アプローチ :TokenCut

TokenCutアルゴリズムは境界を予測するために使用できる。画像内の目立つオブジェクトの位置を示すボックス。図2に示す私たちのアプローチは、ノードはトークンであり、エッジは類似点である。潜在変数に基づく特徴を用いたトークン変圧器について。以下では、まず簡単にセクション3.1.1のビジョントランスフォーマーと正規化カット3.1.2節で、我々の解決策と

実装の詳細についてはセクション3.2を参照してください。

3.1 背景

3.1.1 ビジョントランスフォーマー

H × Wのサイズの画像が与えられた場合、視覚変換器は（ViT） [18]は、解像度K×Kの重複しない2D画像パッチを入力として取り、パッチの数は N = HW/K2 . 各パッチはトークンとして表現され、数値特徴のベクトルによって記述され、埋め込み。クラスとして表される追加の学習可能なトークンCLSトークンは、パッチセット全体の集約情報を表すために使用されます。このCLSトークンと

パッチトークンのセットは、「プレノルム」層正規化[2]を備えた標準的なトランスフォーマーネットワークに供給されます。

ビジョントランスフォーマーは複数の層で構成されています
フィードフォワードネットワークと自己注意のための複数の注意ヘッドを備えたエンコーダのスキップと並列接続
接続。教師なしオブジェクト発見タスクでは、
自己教師あり学習で訓練されたビジョントランスフォーマーを使用する
DINO [6]を用いた学習と潜在変数の抽出
最終層を、提案の入力特徴として
方法。

3.1.2 正規化カット（Ncut）

グラフ分割。Ncut [43]はグラフを分割するために使用できる。
グラフを2つの互いに素な集合AとBに分割する。この手法では、Ncut
エネルギーを最小化するようにグラフを分割する[43]。

$$\text{Nカット}(A, B) = \frac{\text{タクシー)}}{C(A,V)} + \frac{\text{タクシー)}}{C(B,V)} \tag{1}$$

ここでCは2つの
集合C(A, B) = Ei,jであり、C(A, V)は
vi∈A、vj∈Bは
Aのノードからグラフ内のすべてのノードへの合計接続です。

ShiとMalik [43]が示したように、式1の最適化問題は次式と同等である。

$$\min_x \text{Nカット}(x) = \min_y \frac{y^T (D - E)y}{y^T Dy}。 \tag{2}$$

条件y∈{1, −b}
Dは、 di =の対角行列である。
北、bはyを満たす $y^T D1 = 0$です。
j 対角線上のEi,j 。

緩和された制約条件を持つNcut解。z = $D^{\frac{1}{2}}y$. 式2は次のように書き直すことができます。

$$\text{ミンツ} \frac{z^T D^{-\frac{1}{2}}(D - E)D^{-\frac{1}{2}}z}{z^T z}。 \tag{3}$$

[43]に示されているように、式3の定式化は
レイリー商[52]に等しく、これは次のように解ける。
$D^{-\frac{1}{2}}(D-E)D^{-\frac{1}{2}}z = \lambda z$、ここでD − Eはラプラ
アン行列であり、半正定値行列であることが知られている[38]。
したがってz0 = $D^{\frac{1}{2}}1$ は、
最小の固有値λ = 0である。レイリー商[52]によれば、 2番目に小さい固有ベクトルz1は最小の固有ベクトル(z0)に垂直であり、式3のエネルギーを最小化するために使用できる。

$$z1 = \underset{z \perp z0}{\operatorname{argmin}} \frac{z^T D^{-\frac{1}{2}}(D - E)D^{-\frac{1}{2}}z}{z^T z}。$$

z = Dとすると $D^{\frac{1}{2}}$歳、

$$y1 = \underset{y^T D1=0}{\operatorname{argmin}} \frac{y^T (D - E)y}{y^T Dy}。$$

したがって、一般化された
固有値（D − E)y = λDyは実数値解である。
Ncut [43]問題へ。

3.2 トークンカットアルゴリズム

グラフ構築。本手法では、 3.1.1節で説明した視覚変換を用いてベクトルを生成する。
各K×K画像パッチの特徴。完全に接続された
パッチの無向グラフG = (V, E)が構築される
ここで、各Vは特徴ベクトルを持つパッチを表す。
$\{v_i\}_{i=1}^{北}$, そして各パッチは隣接するパッチとリンクされている。
ラベル付けされたエッジ、E。エッジラベルは類似度スコアを表す。
Sの特徴ベクトルのコサイン類似度に基づいて
2つのパッチ。

$$E_{i,j} = \begin{cases} 1、S(v_i, v_j) \geq \tau \text{の場合} \\ 、\text{それ以外} \end{cases}、 \tag{4}$$

ここでτはハイパーパラメータであり、小さな値である。
1e − 5 はグラフが完全に連結であり、Sが
特徴間のコサイン類似度。空間情報は、Transformerの位置エンコーディング
によって暗黙的に特徴に含まれていることに注意してください。

$$S(v_i, v_j) = \frac{v_i v_j}{\|v_i\|_2 \|v_j\|_2}。 \tag{5}$$

3.1.2節で述べたように、Ncutアルゴリズムを適用する。
構築されたグラフG上で、2番目に小さい
一般化固有系の固有ベクトルは、
潜在的なオブジェクトの注目マップとして捉えることができます。この
注目マップの可視化については第4節で示します。

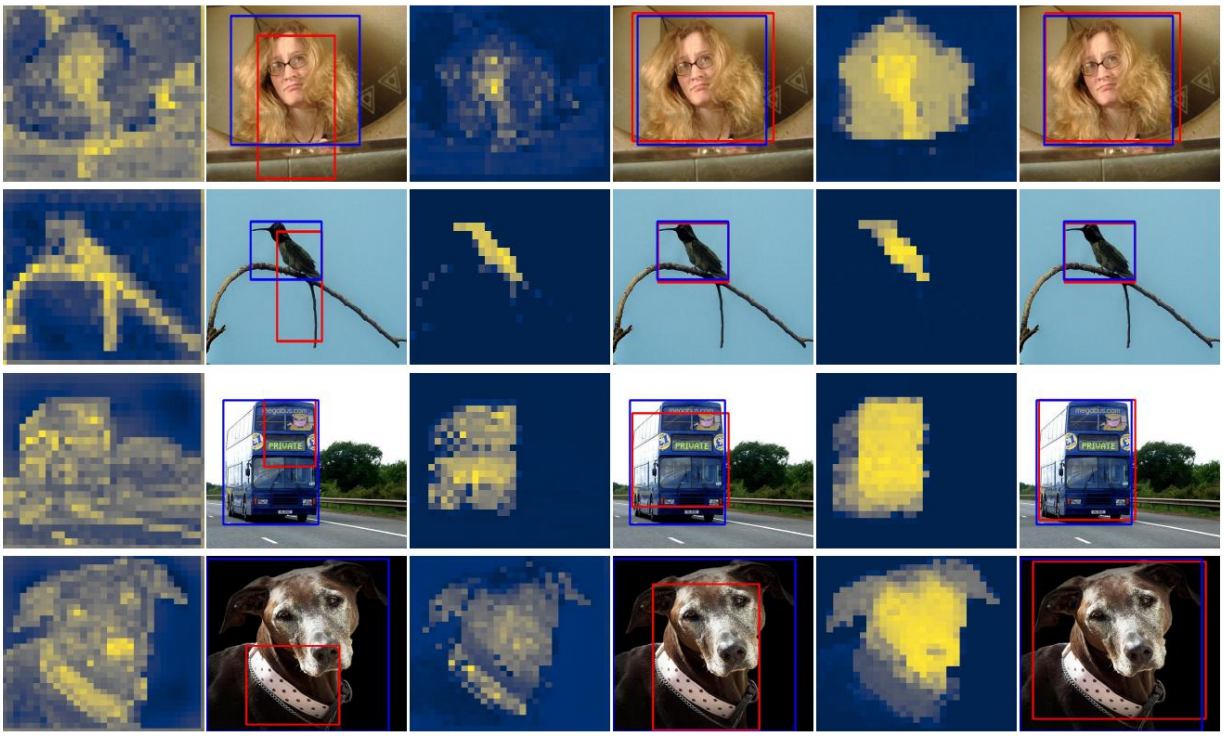
TokenCutで顕著な物体を発見する。画像には少なくとも1つの物体があり、

オブジェクトが前景領域を占めます。
画像から前景のオブジェクトを分割するには、
3つの問題を解決する必要があります。i)
グラフを2つのサブグラフに分割し、ii) グラフの2分割が与えられた場合、どの分割が
前景を表す。iii) 前景に複数の連結成分を検出する場合、

最も顕著なオブジェクトを特定します。

最初の問題については、初期の実験では
投影の単純な平均値を使用して、
類似性を決定するための2番目に小さい固有ベクトル
グラフy1をカットするための値= $\frac{1}{|V|}\sum_{i=1}^{北} y1_i$. 正式には、
A = {vi |y $_1^{北}$ ≤ $\overline{y1}$}かつ B = {vi |y $_1^{北}$ > $\overline{y1}$}である。
これを従来のクラスタリングアルゴリズムと比較すると
K平均法とEM法を用いて2番目に小さい固有値をクラスタ化する
ベクトルを2つの領域に分割します。比較は
補足資料表7によると、平均
一般的に、より良い結果が得られます。

2つ目の問題では、前景には
顕著なオブジェクトであり、あまり関係がないと想定されています



(a)ディーノCLS トークンの注目 (b) ディノ 検出 (c) LOST インバース 度数注意 (d) 失われた 検出 (e) 私たちのアイゲン 注意 (f)当社の 検出

図3: VOC12を用いた教師なし単一物体検出の視覚的結果。 (a)は、検出に用いられるDINO [6]のCLSトークンの注目度を示している(b)。LOST [45]は、主に逆次数マップ(c)を用いて検出を行っている(d)。本手法では、固有ベクトルを(e)に、検出結果を(f)に示す。青と赤の境界ボックスは、それぞれ正解境界ボックスと予測境界ボックスを示している。

グラフ全体。直感的に、 viが前景に属し、 vjが背景トークンである場合、 di < dj となります。したがって、前景オブジェクトの固有ベクトルは、背景の固有ベクトルよりも絶対値が大きくなければなりません。最大絶対値vmaxを用いて、前景パーティションと最も目立つオブジェクトを選択します。vmaxを含むパーティションが前景となります。グラフには明示的な空間制約がないため、前景には複数の連結領域が含まれる可能性があります。最終的なオブジェクト領域として、最大絶対値vmaxを含む前景に存在する最大の連結コンポーネントを選択します。

要約すると、TokenCut アルゴリズムは次の手順で構成されます。

1. 画像が与えられたら、式4と式5に従ってグラフG = (V, E)を構築します。
2. 2番目に小さい固有値y1に関連付けられた固有ベクトルについて、一般化固有系(D − E)y = λDyを解きます。
3. y1の平均を使用して二分割を計算する: A = {vi |y ≤ y1}およびB = {vi |y > y1}
4. y1の最大絶対値に関連付けられた最大の接続コンポーネントを見つけます。

実装の詳細。実験では、パッチの特徴を抽出するために自己蒸留損失（DINO） [6]でトレーニングされたViT-S/16モデル[18]を使用します。最終層の主要な特徴を入力特徴vとして使用します。さまざまな特徴のアブレーションと自己教師学習でトレーニングされたトランスフォーマーは、補足資料の表5に記載されています。すべてのデータセットでτ = 0.2に設定し、τへの依存性はセクション4.4で提供されています。実行時間に関しては、最適化されていない実装では、単一のGPU QUADRO RTX 8000で解像度480×480の単一画像の境界ボックスを検出するのに約0.32秒かかります。

4つの実験

我々は、教師なし単一物体検出、弱教師物体検出、教師なし顕著性検出の3つのタスクで本手法を評価する。教師なし単一物体検出の結果は4.1節に示す。弱教師物体検出の結果は4.2節に示す。教師なし顕著性検出の結果は4.3節に示す。τの分析は4.4節に示す。

表1:教師なし単一物体検出の比較。TokenCutと最先端の物体検出アルゴリズムを比較する。
VOC07 [19]、 VOC12 [20] 、COCO20K [33, 56]データセットにおける発見手法。モデルの性能は
CorLocメトリックを使用して、「画像間類似」とは、モデルがデータセット全体の情報を活用し、
画像間の類似性を利用してオブジェクトを特定します。

方法	インターイメージシミ。	DINO [6]フィーチャリング	VOC07 [19]	VOC12 [20]	COCO20K [33, 56]
選択的検索[45, 51]		-	18.8	20.9	31.1
エッジボックス[45, 79]		-	46.4	46.2	50.5
キムら[30, 45]		-	54.5	55.3	53.6
Zhangeら[45, 74]		-	46.2	61.9	64.0
DDT+ [45, 61]	✓	-	72.1 (↑ 8.1)		
rOSD [45, 56]	✓	-			
LOD [45, 57]	✓	-			
DINOセグメント[6, 45]	✓	ViT-S/16 [18]			
失われた[45]	✓	ViT-S/16 [18]			
トークンカット	✓	ViT-S/16 [18]			
LOD + CAD [45]	✓	-	56.3	61.6	58.3
rOSD + CAD [45]	✓	-	65.7	70.4	
LOST + CAD [45]		[18] 71.4 (↑ 5.7)	75.3 (↑ 4.9)		
トークンカット + CAD [45]					

方法	VOC07 [19]	VOC12 [20]	COCO20K [33, 56]
選択的検索[45, 51]	18.8	20.9	31.1
エッジボックス[45, 79]	46.4	46.2	50.5
キムら[30, 45]	54.5	55.3	53.6
Zhangeら[45, 74]	46.2	61.9	64.0
DDT+ [45, 61]	72.1 (↑ 8.1)		
rOSD [45, 56]			
LOD [45, 57]			
DINOセグメント[6, 45]			
失われた[45]			
トークンカット			
LOD + CAD [45]	56.3	61.6	58.3
rOSD + CAD [45]	65.7	70.4	
LOST + CAD [45]	[18] 71.4 (↑ 5.7)	75.3 (↑ 4.9)	
トークンカット + CAD [45]			

方法	インターイメージシミ。	DINO [6]フィーチャリング	VOC07 [19]	VOC12 [20]	COCO20K [33, 56]
選択的検索[45, 51]		-	18.8	20.9	31.1
エッジボックス[45, 79]		-	46.4	46.2	50.5
キムら[30, 45]		-	54.5	55.3	53.6
Zhangeら[45, 74]		-	46.2	61.9	64.0
DDT+ [45, 61]	✓	-	72.1 (↑ 8.1)		
rOSD [45, 56]	✓	-			
LOD [45, 57]	✓	-			
DINOセグメント[6, 45]	✓	ViT-S/16 [18]			
失われた[45]	✓	ViT-S/16 [18]			
トークンカット	✓	ViT-S/16 [18]			
LOD + CAD [45]	✓	-	56.3	61.6	58.3
rOSD + CAD [45]	✓	-	65.7	70.4	
LOST + CAD [45]		[18] 71.4 (↑ 5.7)	75.3 (↑ 4.9)		
トークンカット + CAD [45]					

+CAD は、「疑似ボックス」ラベルを使用して、第 2 段階のクラスに依存しない検出器をトレーニングすることを示します。

tion 4.4 では、その他のアブレーション研究が補足資料として紹介されます。

4.1 教師なし単一物体検出

評価指標。パフォーマンスは、
CorLoc メトリックは、 [11, 15,
47, 55–57, 61]。CorLocは予測された境界ボックスをカウントする。
予測された境界ボックスと地面の1つとの間の交差和（IoU)スコアが正しいと判断される。
真実の境界ボックスは 0.5 より優れています。

定量的結果。私たちは、以下の点についてアプローチを評価します。
教師なし単一物体検出のための一般的に使用される3つのベンチマーク：
VOC07 [19] 、VOC12 [20] 、および
COCO20K [33, 56]。定性的な結果は表1に示す。CorLocスコアは、

これまでの最先端の単一オブジェクトとの比較
発見方法[30, 45, 51, 56, 57, 61, 74, 79]
VOC07、VOC12、COCO20Kデータセット。これらの手法は、大きく分けて2つのグループに分けられます。
モデルがデータセット全体の情報を活用し、画像間の類似性を探索するかどうか。
地域比較の複雑さが2乗であるため
画像間では、画像間の類似性を持つモデルが
一般的に、大規模なデータセットに拡張するのは困難です。選択的な
検索[51]、エッジボックス[79]、 LOST [45] 、TokenCut
画像間の類似性を必要としないため、
より効率的です。表に示されているように、TokenCutはすべてのデータセットにおいて、一貫してすべての従来の手法よりも優れています。
大きな差です。具体的には、TokenCutはVOC07、VOC12において、最先端技術を6.9%、8.1%、8.1%向上させました。
および COCO20K で、それぞれ同じ ViT-S/16 機能を使用しています。

また、第2段階を含む一連の結果もリストします。
パフォーマンスを向上させるための無監督トレーニング戦略、
これはクラス非依存検出（CAD)と呼ばれます。
CADは、最初の段階の単一モデルによって生成されたすべてのボックスに同じ「前景」カテゴリを割り当てることによってトレーニングされます。
オブジェクト発見モデル。表1に示すように、TokenCut +
CADは最先端のものより5.7%、4.9%、
VOC07、VOC12、COCO20kではそれぞれ5.1%です。

定性的な結果。図3では、
DINO-seg [6]、LOST [45] 、Tokencutの場合。それぞれ
この方法では、実行に使用されるヒートマップを視覚化します
物体検出。DINO-segの場合、ヒートマップはCLSトークンに関連付けられた注目マップです。LOSTの場合、
検出は主に逆次数マップに基づいている
1
TokenCutの場合、2番目に小さい固有値を表示します。
ベクトル。視覚的な結果は、Tokencutが
目立つオブジェクトの高品質なセグメンテーションを抽出します。
DINO-segやLOSTと比較すると、TokenCutは
より完全なセグメンテーションを抽出するために、
図3の最初のサンプルと3番目のサンプルでは、
全ての手法が高品質のマップを持っている場合、To-kenCutはオブジェクト上で最も強い強度を持ちます。この現象は図3の最後のサンプルで確認できます。
視覚的な結果は補足資料に記載されています
図7と図8。

4.2 弱教師あり物体位置推定

評価指標。以下の3つの標準指標を報告します。
Top-1 Cls、GT Loc、Top-1 Loc。Top-1 Clsは
画像分類におけるトップ1精度。GT LocはCorLocに似ており、予測された境界ボックスと真の境界ボックスの1つとの間のIoUスコアが0.5を超える場合、予測ボックスは正解とみなされます。Top-1 Locは最も重要な指標であり、

表2:弱教師あり物体位置推定の比較。Top -1 Cls、GT Loc、Top-1 Locを報告します。
CUB [59]とImageNet-1K [14]のデータセットで比較した。比較した最先端の手法は2つのグループに分けられる。
ImageNet-1K 教師あり事前トレーニングと ImageNet-1K 自己教師あり事前トレーニング。

事前学習済みデータセット	方法	バックボーン	CUB [59]、平均（%）			ImageNet-1K [14]、精度（%）	
			トップ1クラス	GTロケーション	トップ 1 ロック	トップ 1 Cls GT ロック	トップ 1 ロック
イメージネット-1K [14] 教師あり事前学習	CAM [77]	グーグルネット[48]	73.8	HaS-32	-	41.1	65	-	43.6
	[46] + [3]	グーグルネット[48]	75.4	ADL [13] + [3]	61.1	47.4	68.9	60.6	44.6
		ResNet50 [24]	75.0	ADL [13]	77.6	59.5	75.8	62.2	49.4
		インセプションV3 [49]	74.6	I2C	-	53.0	72.8	-	48.7
	[76]	InceptionV3 [49]	-	PSOL	72.6	56	73.3	68.5	53.1
	[71] ‡	インセプションV3 [49]	-	SLT-	-	65.5	-	65.2	54.8
	Net [22]	インセプションV3 [49]	76.4		86.5	66.1	78.1	67.6	55.7
イメージネット-1K [14] 自己教師あり事前学習	LOST [45]	ViT-S/16 [18]	79.5		89.7	71.3	77.0	91.8 (↑ 2.1)	72.9 (↑ 1.6)	77.0	60.0	49
	トークンカット	ViT-S/16 [18]	79.5								65.4 (↑ 5.4)	52.3 (↑ 3.3)

最終的な分類結果を得るために 10 個のクロップ拡張を使用します。そして‡ 分類器と検出器を別々に学習します。

表3:教師なしサリエンシー検出の比較TokenCutと最先端の教師なしサリエンシー検出を比較します。
ECSSD [44]、 DUTS [60] 、DUT-OMRON [67]における顕著性検出手法。TokenCutはより良い結果を達成している。
他の競合アプローチと比較します。

方法	ECSSD [44]			ダッツ[60]			DUT-オムロン[67]		
	maxFβ(%)	IoU(%)	累積(%)	最大Fβ(%)	IoU(%)	精度(%)	最大Fβ(%)	IoU（%）	累積(%)
HS [66]	67.3	50.8	84.7	62.4	36.9	82.6	56.1	43.3	84.3
wCtr [78]	68.4	51.7	86.2	61.1	39.2	83.5	54.1	41.6	83.8
WSC [31]	68.3	49.8	85.2	69.7	38.4	86.2	52.3	38.7	86.5
ディープUSPS [37]	58.4	44.0	79.5		30.5	77.3	41.4	30.5	77.9
ビッグビーガン[58]	78.2	67.2	89.9		49.8	87.8	54.9	45.3	85.6
E-BigBiGAN [58]	79.7	68.4	90.6		51.1	88.2	56.3	46.4	86.0
失われた[42, 45]	75.8	65.4	89.5		51.8	87.1	47.3	41.0	79.7
LOST [42, 45]+双対性ソルバー[5]	83.7	72.3	91.6		57.2	88.7	57.8	48.9	81.8
トークンカット	80.3	71.2	91.8	57.6	87.4 (↑ 3.6)	57.2 (↑ 4.9)	93.4 (↑ 1.8)	90.3	60.0	53.3	88.0
TokenCut + バイラテラルソルバー[5]	75.5 (↑ 5.8)	62.4 (↑ 5.2)	91.4 (↑ 2.7)	69.7 (↑ 11.9)	61.8 (↑ 12.9)	89.7 (↑ 7.9)			

分類と検出の両方の測定を考慮している。予測された境界ボックスは、画像のクラスが正しく予測され、

IoUは予測された境界ボックス間の0.5を超えている
そして、グラウンドトゥルースの境界ボックス。

結果 :弱教師あり物体位置推定におけるモデル性能を評価するために、
CUB-200-2011 [59] (CUB) とImangeNet-1k [40] の2つのデータセットを用いた。微調整の詳細については補足資料を参照のこと。

表2では、TokenCutを最先端の弱教師あり物体位置推定手法と比較しています。
CUBとImageNet-1Kデータセット。手法は2つのグループに分けられる：
ImageNet-1Kの教師あり事前学習で初期化されたモデル[3, 13, 22, 46, 71, 76, 77]と
ImageNet-1K自己教師事前学習[45]で初期化されたモデル。

CUBデータセットでは、TokenCutはすべての方法の中で最高のパフォーマンスを達成し、GT Locでは最先端のLOST法を2.1%と1.6%上回りました。
興味深いことに、ImageNet-1Kの自己教師あり事前学習モデルはすべて、教師あり事前学習モデルよりも優れています。これは、

教師あり事前学習は、自己教師あり事前学習よりも、事前学習済みデータセットのより識別的な表現を学習する。
事前訓練により、
CUBなどの下流データセット。これと比較して、自己教師あり事前学習はより一般的な表現を学習できるため、より優れた転移性が得られます。

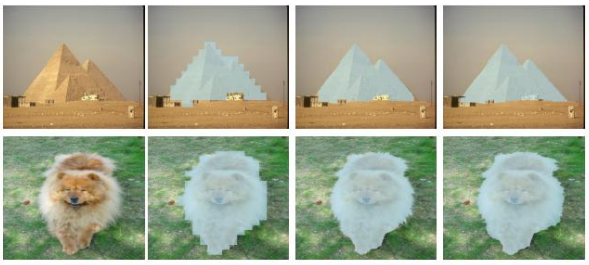
ImageNet-1Kデータセットでは、TokenCutが優れている
GT LocとTop-1 Locでそれぞれ5.4%と4.4%の損失。
ImageNet-1Kの教師あり事前学習モデルと同等の性能を達成した。下流タスクが
ImageNet-1Kの場合、教師あり事前学習は
ImageNet-1K は、データセットに合わせて調整されるため、位置特定タスクを改善する識別機能を提供できます。

表4: τの分析VOC07、VOC12、VOC13、VOC16、VOC18、VOC19、VOC20、VOC21、VOC22、VOC23、VOC24、VOC26、VOC28、VOC29、VOC30、VOC31、VOC32、VOC33、VOC34、VOC35、VOC36、VOC38、VOC39、VOC40、V
COCO20K、および弱教師付きオブジェクトのTop-1 Loc
CUB および ImageNet-1K での検出。

τ	CorLoc トップ1ロケーション				
	VOC07	VOC12	COCO20K	CUB	イメージネット-1K
0	67.4	71.3	56.1	73.0	53.8
0.1	68.6	72.1	58.2	73.2	53.4
0.2	68.8	72.1	58.8	72.9	52.3
0.3	67.7	72.1	58.2	70.8	50.4

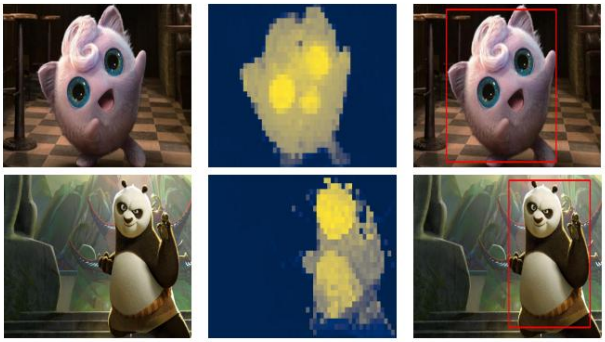
4.3 教師なしサリエンシー検出

評価指標次の 3 つの標準指標を報告します。
F値、IoU、精度。F値は標準的な
顕著性検出における尺度。これはFβ =として計算される。
$\frac{(1+\beta^2)\text{精度}\times\text{再現率}}{\beta^2\text{Precision}+\text{Recall}}$、 適合率と再現率
二値化された予測マスクに基づいて定義され、
グラウンドトゥルースマスク。maxFβは最大値である。
255の均一に分布した二値化閾値。



(a)入力 (b) 当社 (c) 当社 + BS (d) GT

図4:教師なしセグメントの視覚的結果
ECSSD [44]。 (a)に入力画像を示す。TokenCut
検出結果は (b)に示されている。TokenCut + Bilateral
ソルバーの結果は (c) に示されています。(d) は真実値です。



(a)入力 (b) 固有注意 (c) 検出

図5:インターネットから取得した画像の視覚的結果。入力画像、固有注意、

それぞれ(a)(b)(c)で最終検出。

先行研究[42, 58]に従い、一貫性のためにβ = 0.3に設定した。IoU
(Intersection over Union)スコアは次のように計算される。
バイナリ予測マスクとグラウンドトゥルースに基づいて、
閾値は0.5に設定されています。精度は、正しく割り当てられたピクセルの割合を測定します。
オブジェクト/背景。二値化閾値は
マスクの場合は0.5。

結果我々はさらに、無監視サリエンシー検出のための一般的な3つのデータセット、
EC-SSD [44]、 DUTS [60] 、 DUT -OMRON [67]でTokenCutを評価した。

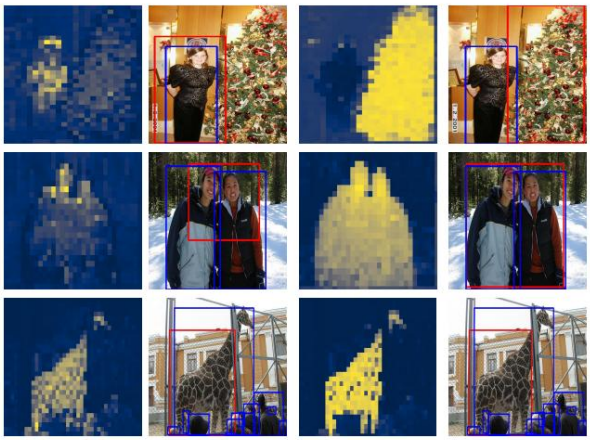
質的結果は表3に示されています。TokenCutは大幅に
これまでの最先端技術を凌駕する性能。二国間
ソルバー[5]はオブジェクトの境界を精緻化し、さらに
TokenCutよりもパフォーマンスが向上し、
図4に示す視覚的な結果からわかる。

4.4 分析と考察

τの分析表4では、式4で定義されたτの分析を示しています。結果は、

τ値の変動は重要ではなく、適切な
閾値はτ = 0.2です。

インターネット画像インターネット上でTokenCutをさらにテストします
画像です。結果は図5に示されています。



(a) 失われた逆アテンド (b) 失われた検出 (c) 私たちのアイゲン注意 (d)私たちの検出

図6: VOC12の故障事例（1行目と2行目）
そしてCOCO （3行目）。LOST [45]は主に
逆次数マップ（a)を用いて検出（b)を実行する。
私たちのアプローチでは、(c)に固有ベクトルを示し、
(d)の検出。青と赤の境界ボックスは
それぞれ真実値と予測された境界ボックスです。

入力画像にノイズの多い背景があっても、私たちのアルゴリズムは正確なアテンションマップを提供してオブジェクトをカバーし、正確なバウンディングボックス予測につなげることができる。
これは、私たちのアプローチの堅牢性を改めて証明しています。

制限事項トークンカットは優れたパフォーマンスを発揮しますが、いくつかの制限事項があります。失敗例
図6に示されている :i) TokenCutは最大の
画像の目立つ部分、望ましくない部分
ii) LOST [45]と同様に、 To-kenCutは、単一の顕著なオブジェクトが

前景。複数の重なり合ったオブジェクトが前景に存在する場合
画像の場合、LOSTと私たちのアプローチはどちらも検出できません
物体の1つ（図6、2行目）。iii) LOSTでも
私たちのアプローチは、オクルージョンを処理できます（図6、3行目）。

5 結論

我々は、教師なしオブジェクト発見のためのシンプルだが効果的なアプローチであるTokenCutを導入しました。TokenCutは
トランスフォーマーを用いた自己教師学習で構築する
グラフでは、ノードはパッチ、エッジはパッチ間の類似性を表します。顕著なオブジェクトは
Ncutを使えば直接検出して区切ることができます。
このアプローチを教師なし単一物体検出、弱教師物体検出、教師なし顕著性検出で評価し、それが

従来のアプローチに比べて大幅に改善されました。
結果は、自己教師あり変換が、おそらく

さまざまなコンピューター ビジョンの問題に使用されます。

謝辞この研究は、MIAI Multidisciplinary AI Instituteの支援を受けて行われました。

大学 グルノーブル アルプ (MIAI@グルノーブル アルプ - ANR-19-P3IA-0003）、およびEU H2020 ICT48によって EU契約番号952026に基づくプロジェクトHumane AI Net。

参考文献

[1] ハメド•H•アグダム、アベル•ゴンザレス=ガルシア、ジューント•ファン•デ Weijer、Antonio M Lopez. 深層検出ニューラルネットワークのための能動学習. ICCV, 2019. 1

[2] ジミー•レイ•バ、ジェイミー•ライアン•キロス、ジェフリー•E•ヒントン レイヤー正規化。arXiv、2016年4月

[3] ペ•ウォノ、ノ•ジュンヒョク、キム•ゴンヒ。再考する 弱教師付き物体局所化のためのクラス活性化マッピング。ECCV, 2020. 7, 16

[4] Hangbo Bao、Li Dong、Furu Wei. Beit: 画像変換機のBert事前学習. arXiv, 2021. 2

[5] ジョナサン•T•バロンとベン•プール。高速双対問題ソルバー。 ECCV、2016年2、7、8頁 ， ，

[6] マティルド•カロン、ウーゴ•トゥヴロン、イシャン•ミスラ、エルベ•ジェグー、ジュリアン•マイラル、ピョートル•ボジャノフスキー、アルマン•ジューラン。 自己教師あり視覚変換における新たな特性。ICCV、2021年。1 , 2, 3, 4, 5, 6, 12, 13, 14

[7] アディティヤ•チャットパディ、アニルバン•サーカール、プランティック•ハウラダー、 Vineeth N Balasubramanian. Grad-cam++: 深層畳み込みネットワークのための 一般化された勾配ベースの視覚的説明. WACV, 2018. 3

[8] 陳信雷、謝彩寧、何開明。経験的な 自己教師あり視覚変換の訓練に関する研究。 ICCV、2021 . 2、12

[9] Yun-Chun Chen、Yen-Yu Lin、Ming-Hsuan Yang、Jia-Bin Huang。表示、一致、分割：弱結合 セマンティックマッチングとオブジェクトコセグメンテーションの教師あり学習。PAMI、2020年。3

[10] ミンミン•チェン、ニロイ•J•ミトラ、シャオレイ•ファン、 フィリップ•HS•トール、シミン•フー。グローバルコントラストに基づく 顕著領域検出。TPAMI、2014年。3

[11] ミンス•チョ、スハ•クァク、コーデリア•シュミット、ジャン ポンセ。教師なし物体検出と位置特定 ワイルド：ボトムアップ領域提案による部分ベースマッチング。CVPR、2015年。2、3、6、16

[12] チェ•ジュンソク、オ•ソンジュン、イ•スンホ、サンヒョク チョン、ゼイネプ•アカタ、シム•ヒョンジョン。評価中 弱教師付き物体位置推定法は正しい。 CVPR、2020. 3

[13] ジュンスク•チェとヒョンジョン•シム。注意に基づく 弱教師付き物体位置推定のためのドロップアウト層。 CVPR、 2019 . 3、7、16

[14] Jia Deng、Wei Dong、Richard Socher、Li-Jia Li、Kai Li、 李菲菲。Imagenet：大規模階層画像 データベース。CVPR 、 2009年。2、7、12、14、15、16

[15] トーマス•デセラース、ボグダン•アレクセ、ヴィットリオ•フェラーリ。 オブジェクトの外観を学習しながら位置を特定する。 ECCV、 2010.6

[16] ジェイコブ•デブリン、ミンウェイ•チャン、ケントン•リー、クリスティーナ Toutanova. Bert: 言語理解のための深層双方向変換器の事前学習。NAACL-HLTでは、 2018年2月

[17] Terrance DeVriesとGraham W Taylor. Cutoutを用いた畳み込みニューラルネットワークの正規化の改良. arXiv、2017年3月

[18] アレクセイ•ドソヴィツキー、ルーカス•ベイヤー、アレクサンダー•コレスニコフ、 ダーク•ワイセンボーン、シャオホア•ザイ、トーマス•ウンターティナー、 モスタファ•デガニ、マティアス•ミンデラー、ゲオルグ•ハイゴールド、 シルヴァン•ゲリー他 1枚の画像は16×16語に相当する： 大規模画像認識のためのトランスフォーマー。ICLRでは、 2020年2月3日、 5日、 6日、 7日、 12日

[19] M. エヴァリンガム、L. ヴァン グール、CKI ウィリアムズ、J. ウィン、 およびA. Zisserman。PASCALビジュアルオブジェクトクラス チャレンジ 2007 (VOC2007) の結果。http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html。 2、6、13

[20] M. エヴァリンガム、L. ヴァン グール、CKI ウィリアムズ、J. ウィン、 およびA. Zisserman。PASCALビジュアルオブジェクトクラス チャレンジ 2012 (VOC2012) の結果。http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html。 2、6

[21] アンドレアス•ガイガー、フィリップ•レンツ、クリストフ•シュティラー、ラケル• Urtasun。ビジョンとロボット工学の融合：Kittiデータセット。 国際ロボティクス研究ジャーナル、2013年1月

[22] Guangyu Guo、Junwei Han、Fang Wan、Dingwen Zhang. 弱教師あり学習における物体位置推定の学習耐性強化. CVPR, 2021. 3, 7

[23] Kaiming He、Xinlei Chen、Saining Xie、Yanghao Li、Pi-otr Dollar、Ross Girshick. マスク付きオートエンコーダはスケーラブルな視覚学習器である. arXiv, 2015. 2

[24] 何開明、張祥宇、任少青、孫建。 画像認識のための深層残差学習。CVPRでは、 2016年7月

[25] ジェフリー•ヒントン、オリオール•ヴィニャルズ、ジェフ•ディーン。蒸留 ニューラルネットワークにおける知識。arXiv、2015年。2

[26] Kuang-Jui Hsu、Yen-Yu Lin、Yung-Yu Chuang、他。教師なしオブジェクトの 共同セグメンテーションのための共同注意 cnn。で IJCAI、2018. 3

[27] 江淮祖、王景東、袁世建、楊呉、 鄭南寧さんと李詩鵬さん。顕著な物体の検出: 識別的な地域的特徴統合アプローチ。 CVPR、2013. 3

[28] Armand Joulin、Francis Bach、Jean Ponce。画像コセグメンテーションのための識別的クラスタリング。CVPR、 2010年3月

[29] アルマン•ジュラン、フランシス•バッハ、ジャン•ポンセ。多クラス 共セグメンテーション。CVPR、2012年。3

[30] Gunhee KimとAntonio Torralba。反復リンク分析を用いた関心領域の教師なし検出。 NeurIPS, 2009年6月

[31] ニアニイ•リー、ビリン•スン、ジンイー•ユー。重み付けされたスパース 顕著性検出のためのコーディングフレームワーク。CVPR、2015年。 3、7

[32] Zhaowen Li、Zhiyang Chen、Fan Yang、Wei Li、Yousong ジュー、チャオヤン、ルイ•デン、リーウェイ•ウー、ルイ•ジャオ、 Ming Tang, et al. Mst: マスク付き自己教師あり変換器 視覚的表現のために。NeurIPS、2021年。2

[33] ツンイー•リン、マイケル•メア、セルジュ•ベロンジー、ジェームズ ， ヘイズ、ピエトロ•ペローナ、デヴァ•ラマナン、ピョートル•ダラー、C•ローレンス •ジトニック。 Microsoft coco: 共通オブジェクト 文脈の中で。ECCV 、 2014年。1、2、6、13、14、16

[34] イェンチェン•リウ、チーヤオ•マー、ジジャン•ホー、チアウェン•クオ、 Kan Chen、Peizhao Zhang、Bichen Wu、Zsolt Kira、 Peter Vajda。半教師付きオブジェクトに対する偏りのない教師 検出。ICLR、2021年1月

[35] Weizeng Lu、Xi Jia、Weicheng Xie、Linlin Shen、Yicong 周、そしてジンミン•ドゥアン。弱く制約された幾何学 教師あり物体位置推定。ECCV、2020年。3

[36] ジンジエ•マイ、メン•ヤン、ウェンフェン•ルオ「統合学習の消去：シンプルだが 効果的なアプローチ」 弱教師付き物体位置推定。CVPR、2020年。3

[37] タム•グエン、マクシミリアン•ダックス、チャイタニヤ•クマール•マムマディ、ヌン ゴ、ティ•ホアイ•フオン•グエン、ゾンユ ルー、トーマス•ブロックス。Deepusps: ディープロバストアンスーパー

自己教師による顕著性予測。NeurIPSでは、
2019年3月7日

[38] Alex Pothen、Horst D Simon、Kang-Pu Liou。グラフの固有ベクトルによる疎行列の分割。SIAM
行列解析と応用ジャーナル、1990年4月

[39] Zhongzheng Ren、Zhiding Yu、Xiaodong Yang、Ming-Yu Liu、Yong Jae Lee、Alexander G Schwing、Jan
Kautz. インスタンス認識、コンテキスト重視、メモリ効率に優れた弱教師あり物体検出。CVPRでは、
2020年1月

[40] オルガ・ルサコフスキー、ジア・デン、ハオスー、ジョナサン・クラウス、
サンジーブ・サシーシュ、ショーン・マー、ジヘン・ファン、アンドレイ
Karpathy、Aditya Khosla、Michael Bernstein他「Ima-genet大規模視覚認識チャレンジ」IJCV、2015年。
7

[41] ランプラーサート R セルバラジュ、マイケル コグズウェル、アビシェク
ダス、ラーマクリシュナ ヴェダンタム、デヴィ パリク、ドゥルブ バトラ。
Grad-cam: ディープネットワークからの視覚的な説明
勾配ベースのローカリゼーションによる。ICCV、2017年。3

[42] Xi Shen、Alexei A Efros、Armand Joulin、Math-ieu Aubry。
セグメントごとの共同セグメンテーションの学習
検索と発見のためのスワッピング。arXivプレプリント
arXiv:2110.15904, 2021年7月8日, 16

[43] Jianbo ShiとJitendra Malik。正規化されたカットと画像
セグメンテーション。TPAMI 、 2000年。2、3、4

[44] 石建平、燕瓊、李徐、佳雅。階層的
拡張CSSDにおける画像の顕著性検出。TPAMI、2015年。
2、7、8、16、17

[45] オリアンヌ・シメオニ、ジル・ピュイ、ユイ・V・ヴォ、シモン・ロブリン、スピロス・ジダリス、
アンドレイ・ブルサック、パトリック・ペレス、ルノー・マーレット、ジャン・ポンセ。自己教師ありトランスフォーマーを使用し、ラベルを使用せずにオブジェクトをローカライズします。 2021 年の BMVC で。
1、2、3、5、6、7、8、12、13、14、15、16

[46] クリシュナ・クマール・シンとヨン・ジェ・リー。かくれんぼ：
弱教師あり学習においてネットワークを細心の注意を払うように強制する
物体と動作の局所化。ICCV、2017年。3、7、16

[47] パルティパン・シヴァ、クリス・ラッセル、タオ・シャン、ルルド
Agapito. 画像の向こう側を見る：物体の顕著性と検出のための教師なし学習. CVPR, 2013. 6

[48] クリスチャン・セゲディ、ウェイ・リウ、ヤンチン・ジア、ピエール・セルマネ、スコット
・リード、ドラゴミル・アンゲロフ、ドゥミトル・エルハン、
ヴィンセント・ヴァンホーク、アンドリュー・ラビノビッチ。
畳み込みによってさらに深くなる。CVPR, 2015. 7, 16

[49] クリスチャン・セゲディ、ヴィンセント・ヴァンフーク、セルゲイ・イオッフェ、ジョン
Shlens、Zbigniew Wojna. コンピュータビジョンのためのインセプションアーキテクチャの再考. CVPR, 2016. 7

[50] ケビン・タン、アルマン・ジュラン、リー・ジア・リー、リー・フェイフェイ。「実世界画像における共局在」CVPR、2014年。3

[51] ジャスパー・RR・ウイリングス、コエン・EA・ヴァン・デ・サンデ、テオ・ゲーヴァース、
アーノルド・W・M・スメルダーズ。選択的オブジェクト検索
認識。IJCV、2013年6月

[52] Charles F Van LoanとG Golub. 行列計算。
ジョンズ・ホプキンス大学出版局、1996年。4

[53] アシシュ・ヴァスワニ、ノーム・シェイゼー、ニキ・パーマー、ジェイコブ
ウシュコライト、ライオン・ジョーンズ、エイダン・N・ゴメス、ルカシュ・カイザー、
イリア・ポロスキン。必要なのは注目することだけ。NeurIPSでは、
2017年2月

[54] サラ・ビセンテ、カーステン・ローザー、ウラジミール・コルモゴロフ
オブジェクトのコセグメンテーション。CVPR、2011年。3

[55] Huy V Vo、Francis Bach、Minsu Cho、Kai Han、Yann Le-Cun、Patrick Perez、
Jean Ponce。最適化としての教師なし画像マッチングと物体発見。CVPR、
2019年2月3日、 6日、 16日

[56] Huy V Vo、Patrick Perez、Jean Ponce. 大規模画像コレクションにおける教師なし複数物体検出に向けて. ECCV, 2020. 2, 3, 6, 13, 14, 16

[57] フイ・V・ヴォ、エレナ・シジコワ、コーデリア・シュミット、パトリック
ペレス、ジャン・ポンセ. 大規模教師なし物体発見. arXiv, 2021. 2, 3, 6, 16

[58] アンドレイ・ヴォイノフ、スタニスラフ・モロゾフ、アルチョム・バベンコ。
大規模生成モデルを用いたラベルなしの物体セグメンテーション。ICML、
2021年。3、7、8

[59] C. Wah, S. Branson, P. Welinder, P. Perona, S. Be-longie. Caltech-UCSD Birds-200-2011データセット. 技術報告書, カリフォルニア工科大学, 2011.
2, 7,
12、14、15、16

[60] ワン・リージュン、ルー・フーチュアン、ワン・イーファン、フォン・メンヤン、
Dong Wang、Baocai Yin、Xiang Ruan。画像レベルの教師あり学習による顕著な物体の検出。CVPRでは、
2017年2月、 3月、 7月、 8月、 16月、 18日

[61] ウェイ・シウシェン、チャン・チェンリン、呉建信、春華
シェン、周志華。教師なし物体発見
深層記述子変換によるパターン認識と共局在化。パターン認識、2019年。2、3、
6、16

[62] 魏宜辰、方文、朱王江、孫建。
背景事前分布を用いた測地線の顕著性。ECCVでは、
2012年3月

[63] ビチェン・ウー、フォレスト・イアンドラ、ピーター・H・ジン、クルト・ケウツァー
Squeezedet: 統合型、小型、低消費電力、リアルタイム物体検出のための完全畳み込みニューラルネットワーク
自動運転。CVPRW、2017年。1

[64] グイソン・シア、シャン・バイ、ジャン・ディン、ジェン・ジュー、セルジュ
ベロンジー、ジエボ・ルオ、ミハイ・ダッチュ、マルチェロ・ペリロ、
Liangpei Zhang. Dota: オブジェクト指向の大規模データセット
航空写真における検出。CVPR、2018年。1

[65] シャオ・ジャンション、ジェームス・ヘイズ、クリスタ・A・エヒンガー、オード
オリバとアントニオ・トラルバ。 Sun データベース: 大規模
修道院から動物園までのシーン認識。CVPR、2010年。16

[66] 瓊燕、李徐、建平史、佳雅。階層的
顕著性検出。CVPR、2013年3月7日

[67] チュアン・ヤン、リーヘ・チャン、フーチュアン・ルー、シャン・ルアン、
ミン・シュアン・ヤン。グラフベースによる顕著性検出
多様体ランキング。CVPR、 2013年。2、3、7、8、16、18

[68] Yuan Yuan、Xiaodan Liang、Xiaolong Wang、Dit-Yan Ye-ung、Abhinav Gupta。時系列動的グラフlstm
アクション駆動型ビデオオブジェクト検出のためのICCV、2017年。3

[69] Yuan Yuan、Yueming Lyu、Xi Shen、Ivor Tsang、Dit-Yan Yeung。周辺化された平均注意ネットワーク
弱教師あり学習。ICLR、2019年。3

[70] ユン・サンドゥ、ハン・ドンユン、オ・ソンジュン、サンヒョク
Chun, Junsuk Choe, Youngjoon Yoo. Cutmix: 局所化可能な特徴量を用いた強力な分類器の学習のための正則化戦略. ICCV, 2019. 3

[71] チェンリン・チャン、ユンハオ・カオ、ジャンシン・ウー。弱く監視されたオブジェクトの位置特定に向けたルートを再考します。 CVPRにて、 2020.3、7

[72] チャン・ディンウェン、ハン・ジュンウェイ、ゴン・チェン、ヤン・ミンシュアン。弱く
監視されたオブジェクトの位置特定と
検出：概要。IEEEパターン分析論文集
および機械知能、2021年3月

[73] ジン・チャン、トン・チャン、ユチャオ・ダイ、メルタシュ・ハランディ、
リチャード・ハートリー著「深層教師なしサリエンシー検出：多重ノイズラベリングの観点」CVPRでは、
2018年3月

[74] ランシェン・チャン、ヤピン・ファン、メンヤン・プー、ジアン
Zhang、Qingji Guan、Qi Zou、Haibin Ling。頻繁なマイニングによる単一のラベルなし画像からの物体検出
マルチスケール特徴を持つアイテムセット。TIP、2020年6月

[75] シャオリン・チャン、ユンチャオ・ウェイ、ジアシー・フォン、イー・ヤン、
Thomas S Huang. 敵対的補完学習
弱教師付き物体位置推定。CVPR、2018年。3

[76] Xiaolin Zhang、Yunchao Wei、Yi Yang。弱教師付き局所化のための画像間通信。
ECCV、2020年。7 [77] Bolei Zhou、Aditya Khosla、Agata Lapedriza、Aude
Oliva、Antonio
Torralba。識別的な局所化のための深層特徴の学習。CVPR、2016年。3、7、16 [78]
Wangjiang Zhu 、Shuang Liang、Yichen Wei、Jian Sun。

堅牢な背景検出からの顕著性の最適化。
[79] C Lawrence Zitnick 、
とPiotr Dollar「エッジボックス：エッジからのオブジェクト提案の検出」ECCV, 2014年
6月

バックボーンの分析。

表5では、異なる変圧器バックボーンのアブレーション研究を示しています。「-S」と「-B」はViT small[6, 18]とViT
それぞれbase[6, 18]アーキテクチャです。「-16」と「-8」はそれぞれパッチサイズ16と8を表します。「MocoV3」は別の
事前学習済みの自己教師ありTransformerモデル[8]。MoCov3の場合、τ値は0.3に設定されていますが、Dinoの場合、最良のτ値は0.2です。
MoCov3の結果はDinoを使用したTokenCutの結果よりわずかに悪いものの、MoCov3は依然として優れていることがわかります。
これまでの最先端技術と比較して、TokenCutは他の自己教師ありトランスフォーマーと組み合わせて使用した場合も同様の結果を提供できることが示されています。
アーキテクチャ。さらに、結果は、パッチサイズ16の方がパッチサイズ8よりも良い結果をもたらすことを示しています。いくつかの洞察
次のようなことが分かります: i) TokenCutは、さまざまなバックボーンにおいてLOSTよりも優れています。ii) LOSTは、ヒューリスティックなシード拡張戦略に依存しているため、
バックボーンの種類によってパフォーマンスは大きく異なります。しかし、私たちのアプローチはより堅牢です。

表5:異なるバックボーンの分析。VOC07における教師なし単一物体検出のためのCorLocを報告する。
VOC12、COCO20K。

方法	バックボーン	VOC07	VOC12	ココ20K
LOST [45]	ViT-S/16 [6, 18]	61.9	64.0	50.7
トークンカット	MoCoV3-ViT-S/16 [8, 18]	66.2	66.9	54.5
トークンカット	ViT-S/16 [6, 18]	68.8 (↑ 6.9)	72.1 (↑ 8.1)	58.8 (↑ 8.1)
LOST [45]	ViT-S/8 [6, 18]	55.5	49.5	57.0
トークンカット	ViT-S/8 [6, 18]	67.3 (↑ 11.8)	71.6 (↑ 14.6)	60.7 (↑ 11.2)
LOST [45]	ViT-B/16 [6, 18]	60.1	63.3	50.0
トークンカット	ViT-B/16 [6, 18]	68.8 (↑ 8.7)	72.4 (↑ 9.1)	59.0 (↑ 9.0)

弱教師あり学習による物体位置推定のための異なるバックボーンに関するアブレーション研究を再度実施する。結果を表6に示す。
「-S」と「-B」はそれぞれViTスモール[6, 18]とViTベース[6, 18]アーキテクチャを示す。「-16」と「-8」はパッチを示す。
それぞれサイズ16と8である。我々のアプローチでは、τ = 0.2の結果を報告するが、これはすべてのデータセットで同じである。
ViT-S/8を使用したLOSTでは、LOSTのシード拡張戦略が上位100のパッチに依存しているため、はるかに悪い結果が得られます。
最も低い次数を持つ。パッチの総数が多い場合、提案されたシード拡張戦略では全体をカバーできない。
オブジェクト。私たちのアプローチは、異なるバックボーン上のさまざまなデータセットに対して、より堅牢なパフォーマンスを提供します。

+

表6:弱教師付き物体位置推定のバックボーンの分析。Top -1 Cls、GT Loc、Top-1を報告します。
CUB [59]およびImagenet-1k [14]データセット上の位置。

方法	バックボーン	τ	CUB [50]、平均（%）			ImageNet-1K [11]、精度（%）	
			トップ 1 Cls	GT ロック	トップ 1 ロック	トップ 1 Cls	GT ロック	トップ 1 ロック
ロスト[45]	ViT-S/16 [6, 18]	79.5	89.7	71.3	77.0	72.9 (↑ 1.6)	77.0	60.0	49.0
トークンカット	ViT-S/16 [6, 18]	0.2	79.5	91.8 (↑ 2.1)				65.4 (↑ 5.4)	53.4 (↑ 4.4)
LOST [45]	ViT-S/8 [6, 18]	-	82.3	78.0	64.4	79.4	45.8	38.1
トークンカット	ViT-S/8 [6, 18]	0.2	82.3	89.9 (↑ 11.9)	74.2 (↑ 9.8)	79.4	66.0 (↑ 20.2)	55.0 (↑ 16.9)
ロスト[45]	ViT-B/16 [6, 18]	-	80.3	90.7	72.8	78.3	72.5 (↓ 0.3)	78.3	58.6	48.3
トークンカット	ViT-B/16 [6, 18]	0.2	80.3	90.0 (↓ 0.7)				63.2 (↑ 4.8)	52.3 (↑ 4.0)

B 二分割戦略の分析。

表7 では、2 番目に小さい固有ベクトルを使用してグラフ内のノードを 2 つのグループに分けるさまざまな戦略を検討します。
我々は、平均値（Mean）、期待最大化（EM）、K平均法（K-means)という3つの自然な手法を検討します。
EMおよびK平均法アルゴリズムの実装にはPythonのsklearnライブラリを使用します。EMアルゴリズムでは、反復回数を300に設定し、
各成分はそれぞれ独自の一般共分散行列を持つ。収束閾値は1e-3に設定される。K平均法アルゴリズムでは、
初期化には「k-means++」を使用する。最大反復回数は300回に設定する。収束閾値は1e-4に設定する。
結果は、分岐点として単純な平均値を用いると、ほとんどの場合にうまく機能することを示唆している。また、
最適なNcut(A,B)値に基づいて分岐点を決定する。このアプローチは二次の計算量を必要とするため、
計算。したがって、最終的には廃止されます。

表7:異なる二分割法の分析。教師なし単一物体検出におけるCorLocの結果を報告する。

2 パーティション	VOC07	VOC12	COCO20K
平均	68.8	72.1	58.8
EM	63.0	65.7	59.3
K平均法	67.5	69.2	61.6

C グラフのエッジ重みの分析

このセクションでは、式4に基づくグラフのエッジ重みの定義に関する考察を示します。類似度スコアをエッジ重みとして直接用いる（すなわち、 Eij = S(xi, xj))という手法を検証しました。しかし、負のエッジ値が存在する可能性があり、正規化カットアルゴリズムの仮定に反するため、これは不可能です。そこで、類似度スコアに閾値を設定する（すなわち、 Eij = S(xi, xj) if S(xi, xj) > τ else)という手法も試みました。VOC07データセットでは68.9%、VOC12データセットでは72%という結果となり、これは既に報告されている結果とほぼ一致しています。

D VOC07とCOCO12における教師なし単一物体検出の視覚的結果

VOC07 [19]とCOCO12 [33, 56] における教師なし単一物体検出の視覚的な結果をそれぞれ図7と図8に示す。

各データセットについて、DINO [6]、LOST [45] 、TokenCutの注目マップとバウンディングボックス予測の両方を比較する。DINOの注目マップは、主要特徴量の最終層のCLSトークン注目マップから抽出される。LOSTの注目マップは、LOSTで検出に用いられる逆次数マップである。TokenCutの注目マップは、式2の2番目に小さい固有ベクトルである。これらの結果は、TokenCutが明らかに優れたオブジェクトセグメンテーションを提供していることを示している。

(a)ディーノCLS トークンの注目 | (b) ディノ 検出 | (c) LOST インバース 度数注意 | (d) 失われた 検出 | (e) 私たちのアイゲン 注意 | (f)当社の 検出

図7: VOC07 [19]における教師なし単一物体検出の視覚的結果。 (a) は、検出に使用されたDINO [6]のCLSトークンの注目度を示しています(b)。LOST [45] は、主に逆次数マップ (c) に基づいて検出を実行します (d)。私たちのアプローチでは、固有ベクトルを (e) に、検出結果を (f) に示しています。青と赤の境界ボックスは、それぞれ真の境界ボックスと予測境界ボックスを示しています。

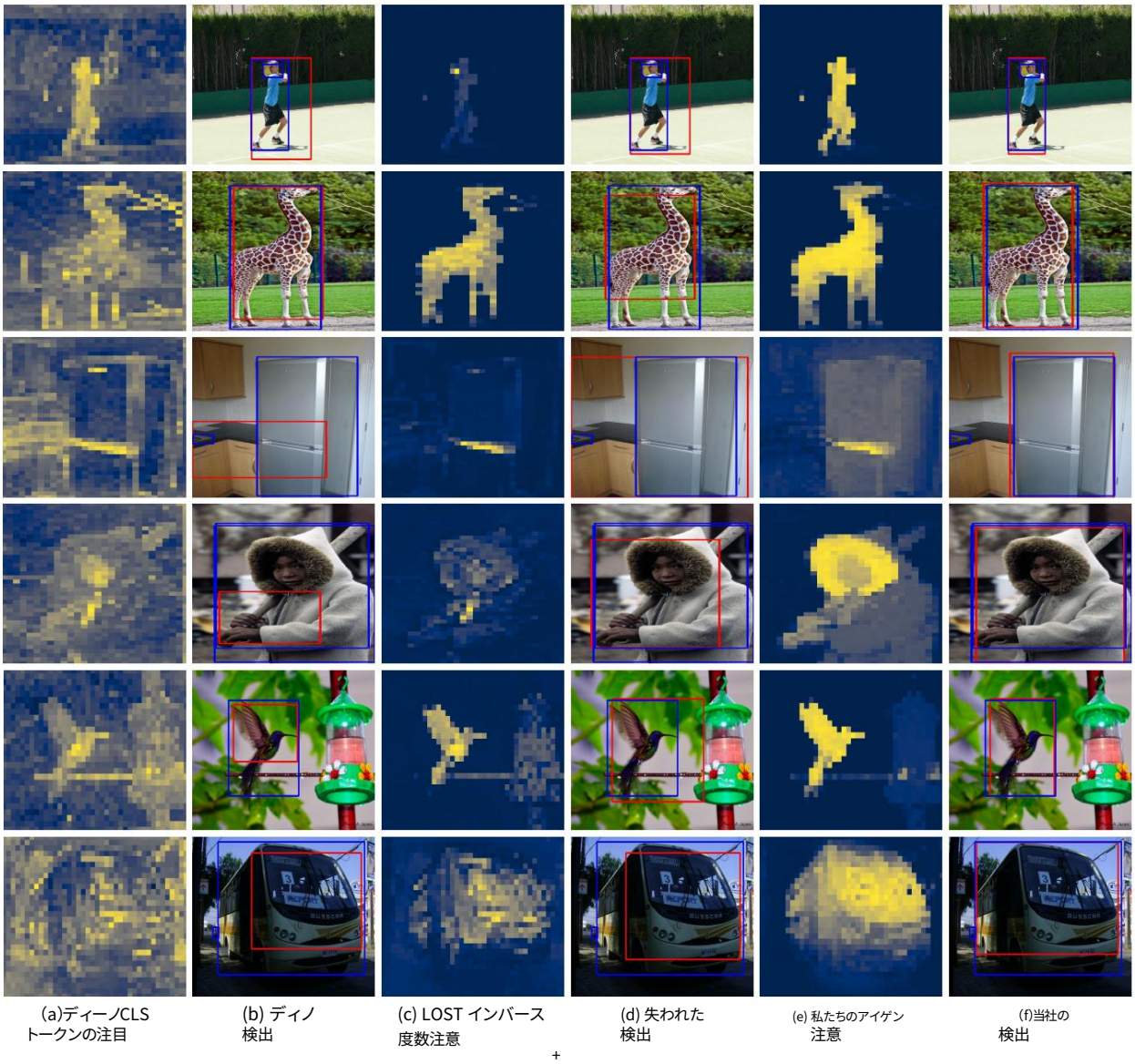


図8: COCO20K [33, 56]を用いた教師なし単一物体検出の視覚的結果。(a) は、検出に使用したDINO [6]におけるCLSトークンの注目度(b)を示している。LOST [45]は主に逆次数マップ(c)を用いて検出を行っている(d)。TokenCutについては、(e) に固有ベクトル、(f) に検出結果を示す。青と赤の境界ボックスは、それぞれ正解境界ボックスと予測境界ボックスを示している。

E 自己教師ありトランスフォーマーの微調整

弱教師あり物体位置推定では、学習済み DINO モデルをバックボーンとして使用し、クラスラベルにのみアクセスできる学習セットで線形分類器を学習します。表2 に示すように、バックボーンの重みを固定し、線形分類器を微調整します。CUB については、SGD オプティマイザーを使用して 1000 エポック学習し、バッチサイズを GPU あたり 256 に設定し、4 つの GPU に分散します。学習率は最初の 50 エポックで線形にウォームアップされ、その後、コサイン学習率スケジューラに従います。学習率は ×5e-4 から 1e-6 に減衰します。重みの減衰は 0.005 に設定されています。ImageNet-1K については、DINO によってリリースされたモデルを使用します。その他の学習設定と詳細については、補足資料をご覧ください。

CUBとImagenet-1kにおける弱教師付き物体位置推定の視覚的結果

CUB [59]とImagenet-1k [14]による弱教師あり物体位置推定の視覚的結果をそれぞれ図9と図10に示す。

各データセットについて、注目マップと境界ボックス予測をLOST [45]と我々のアプローチと比較する。固有ベクトル
TokenCut により、オブジェクトのセグメンテーションが向上し、検出結果が向上します。

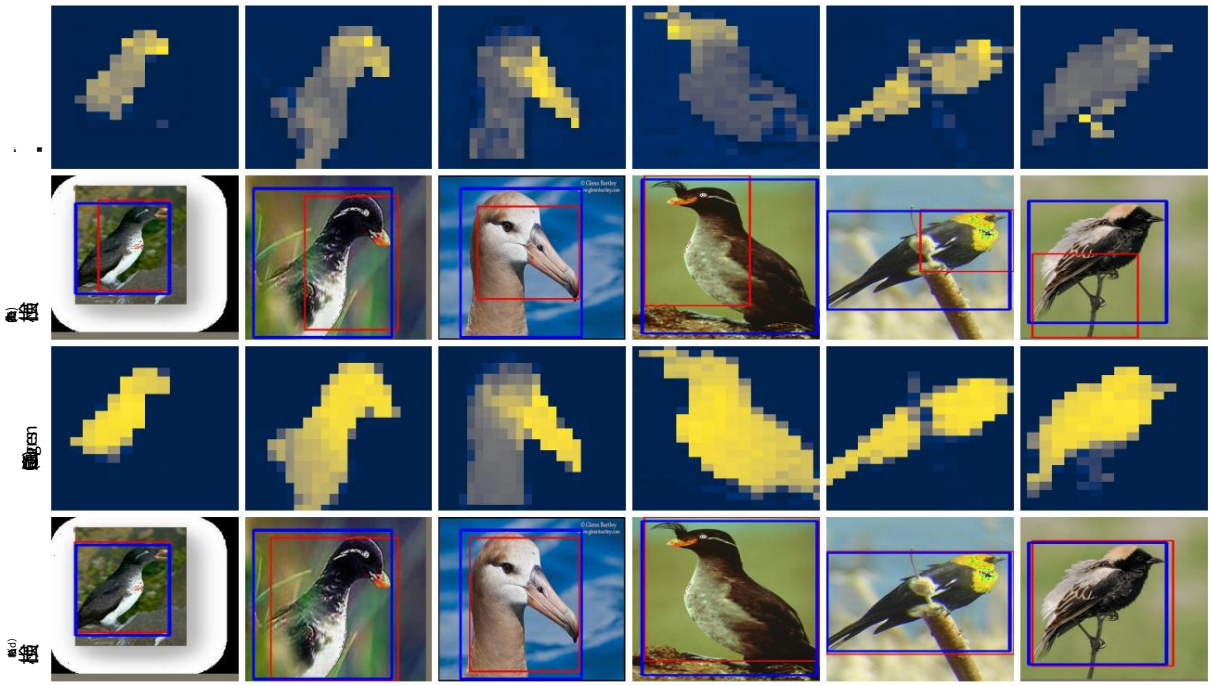


図9: CUB [59] による弱教師付き物体位置推定の視覚的結果。 (a) には、
LOST（b） [45]で検出を行う際に使用される逆次数。TokenCutの場合、固有ベクトルは（c)に示すとおりである。
(d)の検出に使用。青と赤の境界ボックスは、それぞれ真の境界ボックスと予測境界ボックスを示す。
それぞれ。

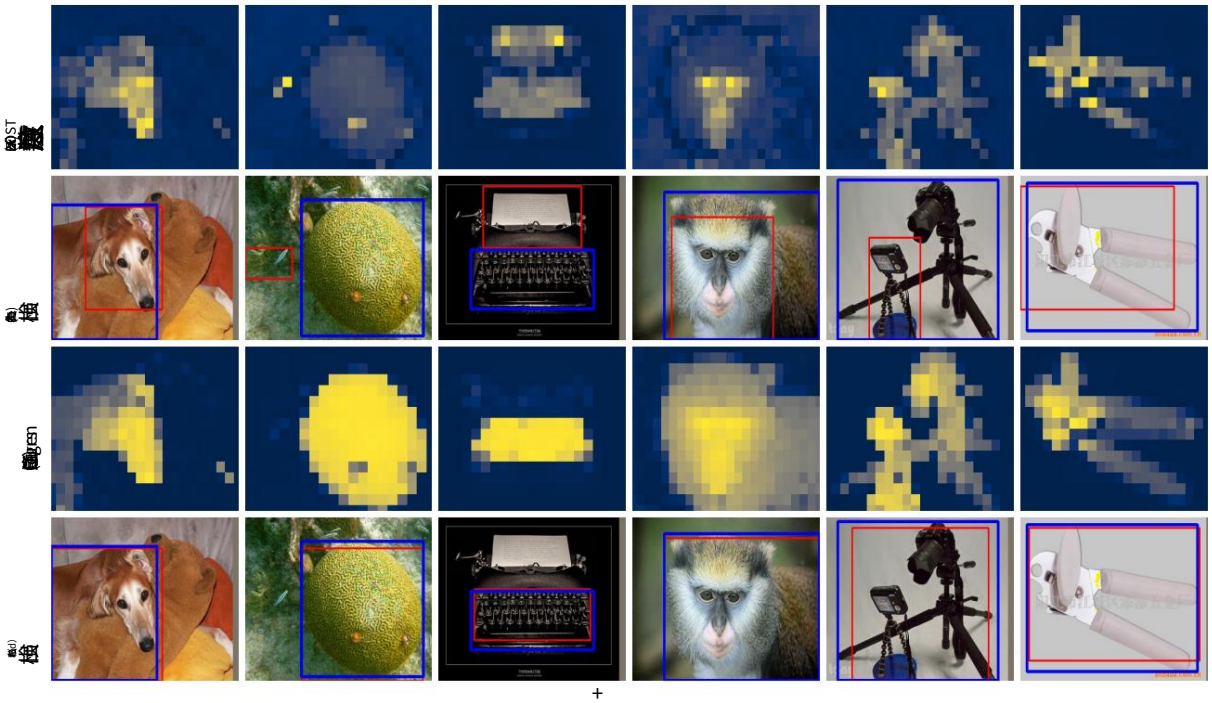


図10: Imagenet-1k [14] における教師なし単一物体検出の視覚的結果。 (a)では、LOST
[45]逆次数マップ。これは検出(b)を実行するために用いられる。TokenCutの場合、固有ベクトルは
(c) と (d) の検出結果。青と赤の境界ボックスは、それぞれ正解と予測境界ボックスを示しています。
それぞれ。

CUBとImagenet-1kにおけるGの失敗事例

図11に、追加の失敗例を示します。これらの失敗例は、3つのカテゴリに分類できます。1) TokenCut
最も目立つオブジェクトに焦点を当てているのに対し、注釈は最初の列と2番目の列に示されている別のオブジェクトを強調表示します。
図11の2) LOSTと同様に、Tokencutは3列目と4列目のような接続されたオブジェクトを区別することができない。
3) 遮蔽がある場合、LOSTも我々のアプローチも、図11の最後の2列のように、物体全体を検出できない。
図11.

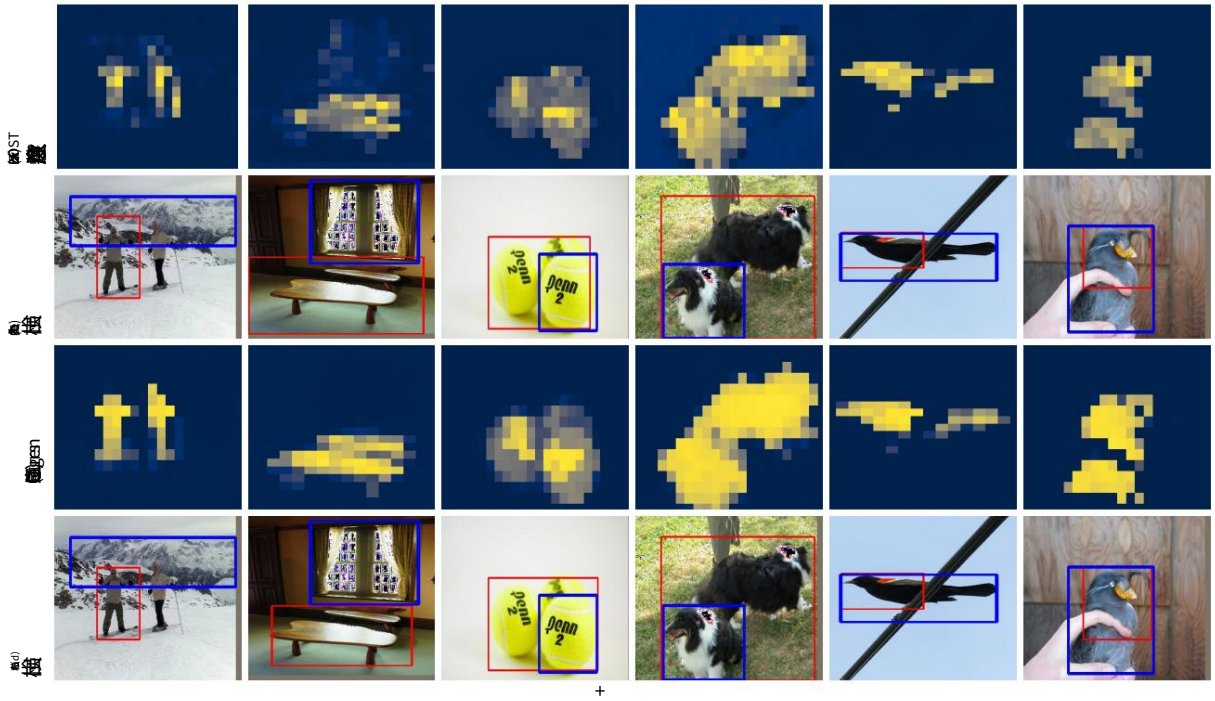


図11: Imagenet-1k [14]とCUB [59] における失敗例。 (a)はLOST [45]の逆次数マップを示し、
これは検出を実行するために使用されます (b)。TokenCutについては、(c)に固有ベクトル、(d)に検出を示します。青
赤い境界ボックスはそれぞれ、真の境界ボックスと予測された境界ボックスを示しています。

Hデータセット

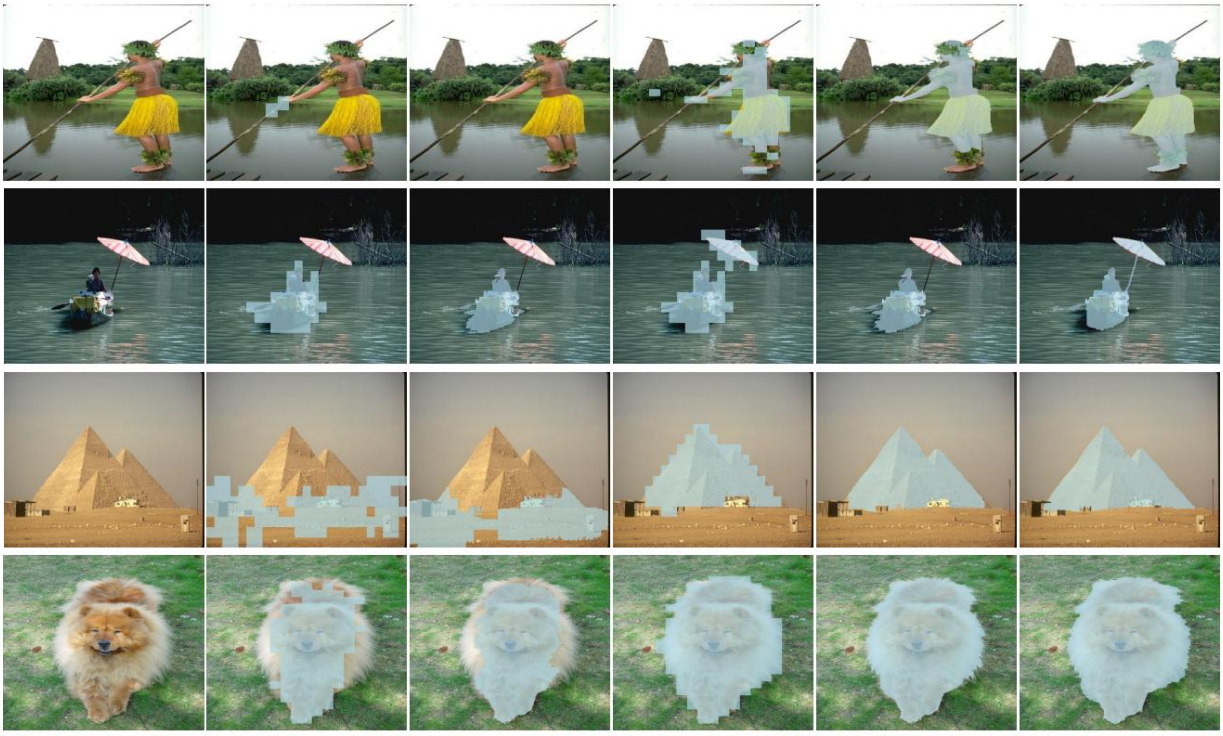
このセクションでは、実験で使用したデータセットの詳細を紹介します。

- VOC07とVOC12はPASCAL-VOC07とPASCAL-VOC12のトレーニングセットと検証セットに対応しています。VOC07と
  VOC12にはそれぞれ5,011枚と11,540枚の画像が含まれており、それぞれ20のカテゴリに属しています。これらは、教師なし物体検出において一般的に評価
  されています[11, 55–57, 61]。
- COCO20Kは、COCO2014データセット[33]からランダムに選ばれた19,817枚の画像から構成されています。 [56]では、
  大規模な評価。
- CUBは200種の鳥類から構成されており、訓練用とテスト用のそれぞれ6,033枚と5,755枚の画像が含まれている。
  弱教師あり学習による物体位置推定の評価に用いられる[3, 13, 46, 48, 77]。
- ImageNet [14]は、画像分類と物体検出のための広く使用されているベンチマークであり、1,000種類の異なる画像から構成されています。
  カテゴリー。訓練セットと検証セットの画像数はそれぞれ130万枚と5万枚です。各画像には
  検出対象となる単一のオブジェクト。トレーニング中はクラスラベルのみが利用可能です。
- ECSSDには、テスト用の複雑なシーンの実際の画像が 1,000 枚含まれています。
- DUTSには10,553枚の学習用画像と5,019枚のテスト画像が含まれています。学習用セットはImageNetの学習用/検証用セットから収集されています。
  テストセットはImageNetテストとSUNデータセット[65]から収集された。先行研究[42]に従って、我々はパフォーマンスを報告する。
  DUTS テスト サブセットについて。
- DUT-OMRON [67]にはテスト用の高品質自然画像が5168枚収録されている。

I ECSSD、DUTS、DUT-OMRONにおける教師なし顕著性検出の視覚的結果

ECSSD [44]、 DUTS [60] 、DUT-OMRON [67]における教師なしサリエンシー検出の視覚的結果を図12、13に示す。
および14です。

各データセットについて、LOSTセグメンテーション、LOST + Bilateral Solver、そして我々のアプローチを比較した。TokenCutはより優れたオブジェクトのセグメンテーション。Bilateral Solver によりパフォーマンスがさらに向上します。

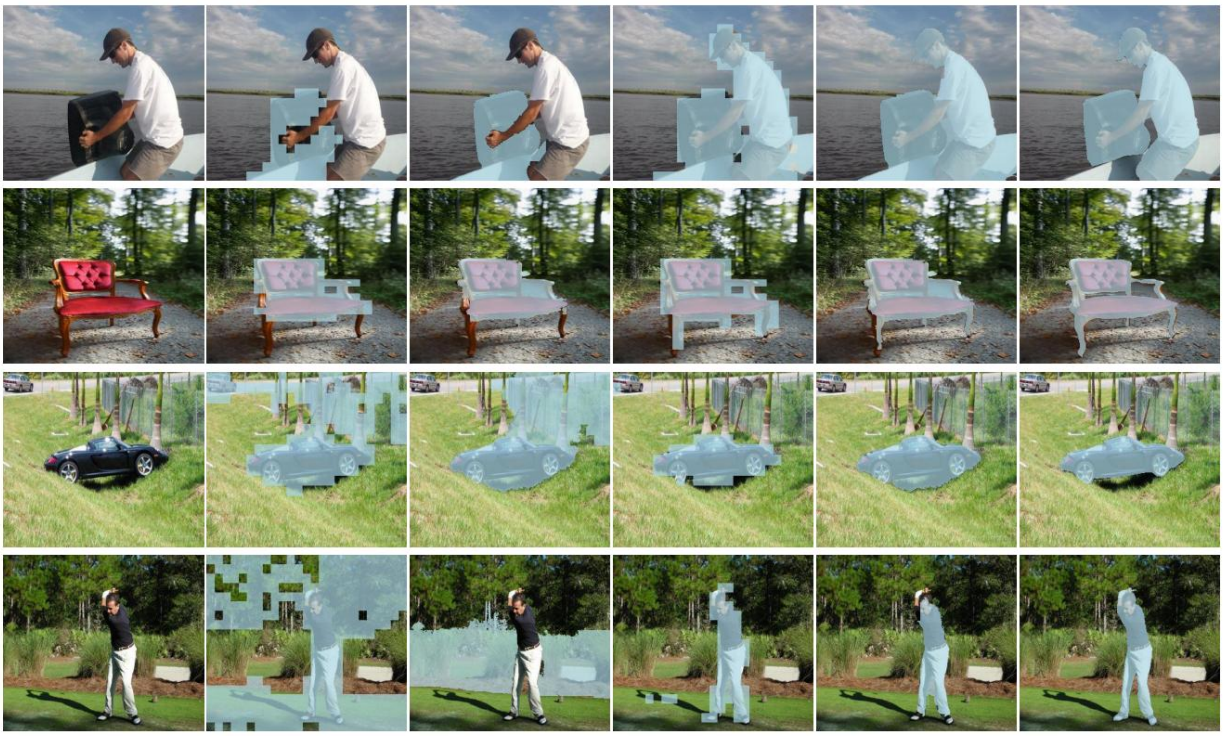


(a)入力 (b) 失われた (c) LOST + BS (e) Ours + BS (d) 私たちの (f) GT

図12: ECSSD上の教師なしセグメントの視覚的結果[44]

(a)入力 (b) LOST (c) LOST + BS (e) Ours + BS (d)私たちの (f) GT

図13: DUTS上の教師なしセグメントの視覚的結果[60]



(a)入力 (b) 失われた (c) LOST + BS (d)私たちの (e) 当社 + BS (f) GT

図14: DUT-OMRON上の教師なしセグメントの視覚的結果[67]