マルチスケールな検出領域を用いた改良型ODAMによる可視化結果の解釈容易性向上

仲井 悠真[1,a)]　平川 翼[1,b)]　山下 隆義[1,c)]　藤吉 弘亘[1,d)]

概要：深層学習モデルによる物体検出は，自動運転や医療画像解析等の分野で幅広く利用されている．特に Transformer を用いた物体検出法はその高い検出精度で注目されているが，モデルの検出結果に対する判断根拠は不明瞭であり，ブラックボックスとされている．この問題に対し，勾配ベースで物体検出結果の判断根拠を可視化する手法として，Object Detector Activation Maps (ODAM) が提案されている．ODAM は検出領域に対するアテンションマップを出力するが，ノイズに敏感であることから，検出した物体以外の領域を強調することがある．そこで本研究では，ODAM の解釈容易性の向上を目的とし，マルチスケールな検出領域を用いた改良型 ODAM を提案する．提案手法では，ODAM に与える検出領域が，可視化結果に大きな影響を及ぼすという性質を利用する．具体的には，検出領域の大きさの変化によって検出物体に注視した可視化結果になるという性質を利用し，異なる拡張率を持つ Bounding Box での可視化結果を平均する．これにより，着目する領域の変動を抑制し，可視化結果の忠実度を維持しつつ解釈容易性を高める．

1. はじめに

深層学習モデルによる物体検出は著しい進歩を遂げており，自動運転や医療画像解析など，幅広い分野で応用されている．これまでの物体検出は CNN[1] に基づく手法が主流であり，R-CNN[2]，You Only Look Once (YOLO) [3]，Single Shot MultiBox Detector (SSD) [4] などが用いられてきた．さらに近年では，Transformer[5] を用いた物体検出手法が，その高い検出精度により大きな注目を集めている．DEtection TRansformer（DETR）[6] は Transformer ベースの物体検出手法の一つであり，End-to-End 学習が可能である．これにより，従来の手法で必要とされていた複雑な後処理を省略することができる．処理の単純化や学習段階の削減が可能となり，従来の手法に比べて効率的な物体検出を可能にしている．

しかし，DETR のような Transformer ベースのモデルは，入力画像のどの特徴に着目して認識しているのか，その判断根拠が不明瞭であり，ブラックボックスとされている．Transformer のような複数のアテンションヘッドを持つモデルの判断根拠を可視化することは困難であり，判断根拠に関しても様々な議論が存在する [7]．このようなモデルのブラックボックス性を解消するためには，判断根拠の可視化が重要であるとされている．モデルの判断根拠を視覚的に示すことにより，推論時のモデルの動作を深く理解し，それによってモデルへの信頼や納得感を高めることができる．具体的には，物体検出モデルが特定の物体を認識した場合，画像内のどの領域がその判断に影響を与えたかを明確にすることが重要であり，誤判定や不確実性の原因を視覚的に特定することが期待できる．さらに，モデルのアーキテクチャや学習プロセスの改善に貢献するためのヒントを得ることも期待でき，より堅牢で正確なモデルの開発や，その改善に不可欠なプロセスである．

従来の可視化手法として，物体検出器の中間特徴マップの勾配情報を利用した Gradient-weighted Class Activation Mapping (Grad-CAM) [8] や Object Detector Activation Maps (ODAM) [9] などがある．これらの勾配ベースの可視化手法はノイズに敏感であり，注視物体以外の領域を誤って強調することがある．そのため，モデルが出力した可視化結果が人間が見たときに理解しやすいかどうかを示す解釈容易性の面では不十分である．解釈容易性が低い可視化結果は，モデルがどの特徴やパターンに基づいて物体を検出しているのかが不明瞭であり，モデルが何を見ているのか識別が困難となる．

そこで本研究では，DETR を用いた物体検出における判

[1] 中部大学
Chubu University
[a)] yuma@mprg.cs.chubu.ac.jp
[b)] hirakawa@mprg.cs.chubu.ac.jp
[c)] takayoshi@isc.chubu.ac.jp
[d)] fujiyoshi@isc.chubu.ac.jp

断根拠の可視化に対する解釈容易性を向上させることを目的とし，検出物体の周辺のコンテキストを考慮した改良型 ODAM を提案する．提案手法では，ODAM に与える検出領域が可視化に大きな影響を与えるという性質に着目し，異なる拡張率の Bounding Box を用いて複数の可視化結果を平均することで，着目する領域の変動を抑制させ，解釈容易性を高める．評価実験により，本手法が従来の ODAM を上回る解釈容易性を持つことを示す．

2. 関連研究

本章では，勾配ベースの判断根拠の可視化手法として Grad-CAM と ODAM について述べる．

2.1 Gradient-weighted Class Activation Mapping

Gradient-weighted Class Activation Mapping (Grad-CAM) [8] は分類タスクにおいてモデルの注目領域を可視化する代表的な手法である．勾配ベースの可視化手法であり，ネットワークの最後の畳み込み層に入力される勾配情報を利用して，予測されたクラスに最も影響を与える画像領域をハイライトする．Grad-CAM は，入力画像がモデルを通過した後に得られる畳み込み層とクラス分類層の出力を利用する．はじめに，クラス分類の出力と各畳み込み層の出力間の関係を捉えるために，誤差逆伝播を用いて勾配を計算することで，クラス分類の出力に対する畳み込み層の出力の重要度を特定する．計算された勾配は，各畳み込み層に対して Global Average Pooling（GAP） を適用することで重み付けされる．GAP は，畳み込み層から得られる勾配の平均値を計算するシンプルな操作であり，この平均値はモデルの意思決定における重要な特徴点となる．得られた重要な特徴は，順伝播での出力に重み付けされ，全畳み込み層にわたって加算されることで，1 枚の活性化マップが生成される．さらに，ReLU 関数を適用することで，負の値を 0 に置き換え，最終的な視覚化結果が得られる．具体的には，はじめにターゲットクラス c に対するネットワークの出力スコア Y^c と，最終畳み込み層の特徴マップ A^k との間の勾配を計算する．この勾配は，式 (1) で示される．

$$\frac{\partial Y^c}{\partial A^k} \tag{1}$$

次に，得られた勾配を GAP によって集約し，クラス c に対する各特徴マップ k の重要度 α_k^c を計算する．この重要度は，式 (2) で計算される．

$$\alpha_k^c = \frac{1}{Z}\sum_i \sum_j \frac{\partial Y^c}{\partial A_{ij}^k} \tag{2}$$

ここで，Z は特徴マップの次元数を指す．最後に，重み付けされた特徴マップの線形結合を行い，クラス c に対する活性化マップ $L^c_{\text{Grad-CAM}}$ を生成する．この活性化マップは，式 (3) によって得られる．

$$L^c_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right) \tag{3}$$

2.2 Object Detector Activation Maps

Object Detector Activation Maps (ODAM) [9] は，インスタンス固有の判断根拠を，勾配ベースで可視化する手法である．物体検出モデルが行った予測の根拠をクラススコアや Bounding Box の座標など，各予測要素に対して重要度をヒートマップとして可視化する．ODAM を用いることで，モデルが各物体をどのように認識し，分類しているか可視化可能である．

ODAM では，与えられた画像 I に対して物体検出モデルで複数の予測を出力し，各予測 p はクラススコア $s(cp)$ と Bounding Box $B(p) = (x_1(p), y_1(p), x_2(p), y_2(p))$ で構成される．特定のインスタンス p に対する予測されたオブジェクト属性スカラー $Y(p)$ を，特徴マップの線形要素単位の加重結合として式 (4) のように示す．

$$Y(p) = \sum_{k,ij} w_{ijk}(p) A_{ijk} \tag{4}$$

ここで，A_{ijk} は特徴マップを表し，$w_{ijk}(p)$ は各ピクセルと各チャンネルの重要性を捉える重みである．次に，インスタンス固有のヒートマップ $H_{ij}(p)$ を生成する．これを式 (5) のように示す．

$$H_{ij}(p) = \sum_k w_{ijk}(p) A_{ijk} \tag{5}$$

最後に，スカラー出力 $Y(p)$ に対応するヒートマップ $H(p)$ は，ピクセルウェイトを用いて得られる．重みマップ $w_k(p)$ は，勾配マップ $\partial Y(p)/\partial A_k$ に局所平滑化処理 Φ を適用することで式 (6) のように定義される．

$$w_k(p) = \Phi\left(\frac{\partial Y(p)}{\partial A_k}\right) \tag{6}$$

ここで，局所平滑化処理 Φ は勾配マップを平滑化し，重要な場所と形状の識別情報を保持しつつ，勾配マップの細かい変動や不規則性，すなわち「ノイズ」と呼ばれる要素を平滑化する役割を果たす．ODAM ではこの処理にガウシアンフィルタを利用しており，この平滑化された重みマップ $w_k(p)$ は，それぞれの特徴マップ A_k に対して要素ごとに乗算され，最終的なヒートマップ $H(p)$ の生成に寄与する．ヒートマップの生成は，式 (7) に示すように，ReLU 関数を使用して行われる．

$$H(p) = \text{ReLU}\left(\sum_k w_k(p) \circ A_k\right) \tag{7}$$

ここで $\circ$ は要素ごとの乗算を意味する．

3. 予備実験

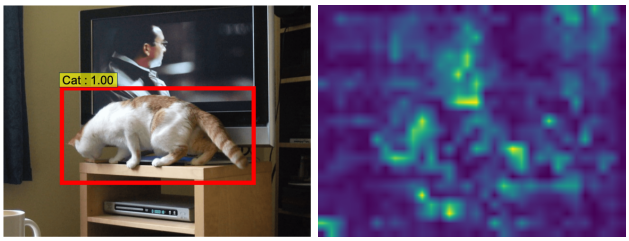
本章では，ODAM を用いて物体検出モデルの判断根拠を可視化し，従来の ODAM の課題を確認するとともに，ODAM に与える検出領域の変更が可視化結果にどのような影響を与えるかを調査する．この予備実験を通じて，ODAM の改善に向けた知見を得ることを目的とする．

3.1 ODAM による判断根拠の可視化

本実験では物体検出モデルとして DETR を採用し，クラス信頼度が 0.7 以上の物体に対して ODAM による可視化を行う．DETR の Backbone には ResNet-50[10] を使用する． 図 1, 図 2 に DETR による物体検出結果に対する判断根拠を ODAM を用いて可視化した結果を示す．図 1(b), 図 2(b) より，ODAM による判断根拠の可視化がインスタンスごとに可能であることがわかる．しかし，ODAM のような勾配ベースの手法はノイズに敏感であり，注視物体以外の領域を誤って強調することがある．図 1(b) に示す例では，検出物体である猫以外の箇所に Attention が拡散しており，モデルがどの部分を判断根拠として物体を猫であると判断したかが不明瞭である．また，ODAM による可視化結果は．図 2(b) に示すように画像の境界領域に帯状のノイズが確認される事がある．このような事象は，モデルの判断根拠を理解する際の妨げとなり，解釈容易性の面で改善の余地があると考えられる．

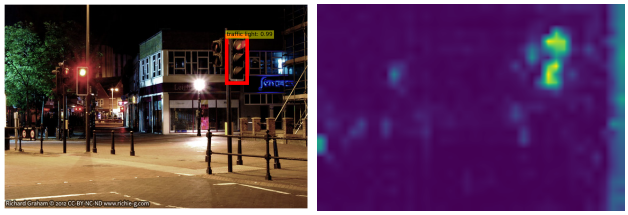
3.2 Bounding Box の大きさ変更に対する可視化

本章では ODAM に入力する Bounding Box の大きさを変更することで，検出物体だけでなく，その周辺のコンテキストを含めた可視化結果が得られる可能性があるという仮説に基づいて実験を行った．ODAM に入力する Bounding Box の範囲を調整した際の可視化結果の変化について調査を行い，調査結果から得られた知見をもとに，ODAM の改良手法を検討する．



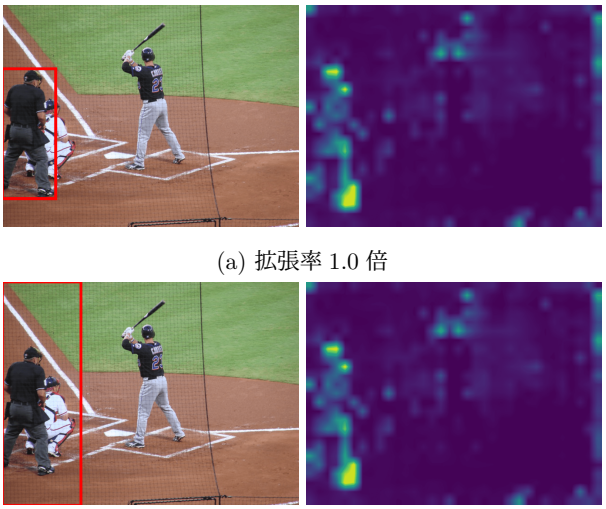
(a) 物体検出結果　　(b) 可視化結果

図 1: 注視物体外への Attention の拡散例



(a) 物体検出結果　　(b) 可視化結果

図 2: 画像境界領域への Attention の集中例



(a) 拡張率 1.0 倍

(b) 拡張率 2.0 倍

図 3: 拡張率による可視化結果の変化

図 3 に画像の一番左側に立つ人物を検出した判断根拠の可視化結果を示す．図 3(a) は Bounding Box 拡張前，図 3(b) は Bounding Box を 2 倍に拡張した際の可視化結果である．図 3 に示すように，ODAM による判断根拠の可視化では，Bounding Box の大きさを拡張することで検出物体の周辺情報を考慮しつつ，より対象物体に着目した可視化結果が得られることがわかる．

さらに, 定量的評価として, COCO データセット [11] に含まれる画像を用いて, Bounding Box の拡張が, ODAM による可視化結果にどのような影響を与えるか，Energy-based Pointing Game (EBPG) [12] を用いて評価する．EBPG の詳細については 5.1.1 節を参照されたい．Bounding Box の拡張率は 1.0 倍から 3.0 倍までの 0.25 倍刻みで変更し，それぞれの倍率での EBPG スコアを評価する. 図 4 に Bounding Box の拡張率を変化させたときのスコアの推移を示す.

図 4 より，1 倍から 3 倍程度の Bounding Box の拡張が，

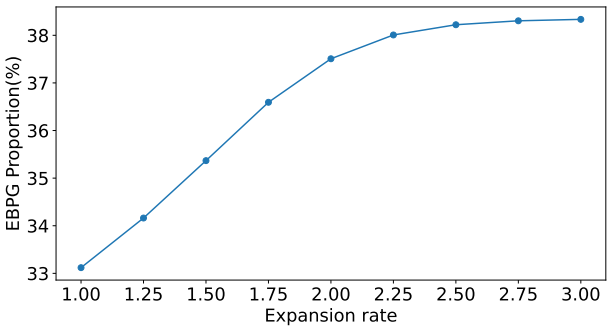


図 4: 拡張率の変化に伴う EBPG の推移

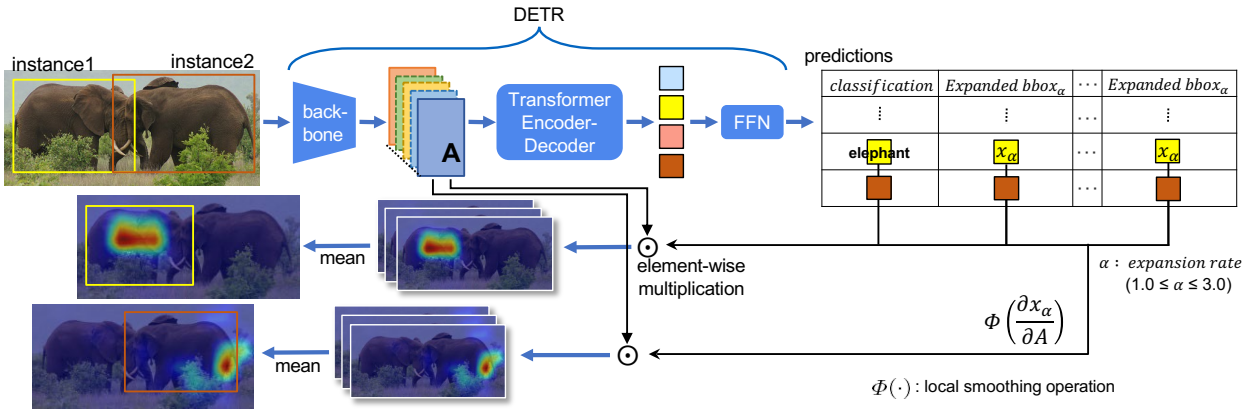


図 5: 提案手法の概要

ODAM による可視化結果の変化に影響を与えており，2.5 倍以降は大きな変化が見られなかったことから，おおよそ 3 倍程度の Bounding Box の拡張が，ODAM による可視化結果に大きく影響を与えることがわかる．Bounding Box 拡張により Attention が検出物体に集中していることが考えられる．

4. 提案手法

可視化結果の解釈容易性の向上を目的として，ODAM の改良手法を提案する．予備実験より，ODAM に与える検出領域が可視化に大きな影響を与えていることがわかった．また，ODAM による可視化では，物体識別能力が高いと，検出物体周辺のコンテキストを考慮しないため，忠実度と物体識別の間にトレードオフの関係があることが報告されている [9]．そこで，提案手法では，異なる拡張率の Bounding Box での可視化結果を平均することで，着目する領域の変動を抑制させ，可視化結果の忠実度を維持しつつ解釈容易性を高める．提案手法の流れを図 5 に示す．

4.1 マルチスケールな検出領域の利用

インスタンス p に対する Bounding Box の左上の座標を $(x_1(p), y_1(p))$，右下の座標を $(x_2(p), y_2(p))$ と定義する．Bounding Box の中心座標 $(cx(p), cy(p))$ は式 (8) のように求められる．

$$cx(p) = \frac{x_1(p) + x_2(p)}{2}, \quad cy(p) = \frac{y_1(p) + y_2(p)}{2} \tag{8}$$

このとき Bounding Box の幅 $w(p)$ と高さ $h(p)$ は式 (9) のように求められる．

$$w(p) = x_2(p) - x_1(p), \quad h(p) = y_2(p) - y_1(p) \tag{9}$$

拡張率を α とすると，拡張された Bounding Box の幅 $w'(p, \alpha)$ と高さ $h'(p, \alpha)$ は式 (10) のようになる．

$$w'(p, \alpha) = w(p) \times \alpha, \quad h'(p, \alpha) = h(p) \times \alpha \tag{10}$$

ここで，予備実験の結果より，拡張率 α の値は 1 から 3 倍が適切であることから，α の範囲は 1 から 3 とする．得られた新たな幅と高さを用いて，拡張後の Bounding Box の座標 $x1'(p, \alpha), y1'(p, \alpha), x2'(p, \alpha), y2'(p, \alpha)$ は式 (11) のように求められる．

$$\begin{aligned} x1'(p, \alpha) &= cx(p) - \frac{w'(p, \alpha)}{2}, \\ y1'(p, \alpha) &= cy(p) - \frac{h'(p, \alpha)}{2}, \\ x2'(p, \alpha) &= cx(p) + \frac{w'(p, \alpha)}{2}, \\ y2'(p, \alpha) &= cy(p) + \frac{h'(p, \alpha)}{2}. \end{aligned} \tag{11}$$

拡張率 α で Bounding Box を拡張するとき，特徴マップ内の座標を $(i\ ,\ j)$，チャンネルを k，検出物体 p に対する特徴マップ A_{ijk} と乗算する重みを $w_{ijk}(p, \alpha)$ としたとき，着目する領域は，式 (12) よりヒートマップ $H_{ij}(p, \alpha)$ をとして求められる．

$$H_{ij}(p, \alpha) = \sum_k w_{ijk}(p, \alpha) \odot A_{ijk} \tag{12}$$

各拡張率 α に対して生成したヒートマップ $H_{ij}(p, \alpha)$ を式 (13) より，平均を求めて最終的なヒートマップ $H'_{ij}(p)$ を獲得する．

$$H'_{ij}(p) = \frac{1}{N} \sum_\alpha H_{ij}(p, \alpha) \tag{13}$$

4.2 DETR への適用

DETR に対して提案手法を適用する際は，まずモデルに画像を入力し，各オブジェクトクエリに対する予測を得る．予測には，クラススコアと Bounding Box の座標が含まれる．次に，特定のオブジェクトクエリ p に対する Bounding Box の座標から，式 (8) を用いて中心座標 $(cx(p), cy(p))$ を求め，式 (9) を用いて Bounding Box の幅 $w(p)$ と高さ $h(p)$ を求める．拡張率 α を 1 から 3 の範囲で変化させながら，式 (10) を用いて拡張後の Bounding Box の幅 $w'(p, \alpha)$ と高さ $h'(p, \alpha)$ を求め，式 (11) を用いて拡張後の Bounding Box の座標 $x1'(p, \alpha), y1'(p, \alpha), x2'(p, \alpha), y2'(p, \alpha)$ を求める．拡張率 α ごとに，式 (12) を用いてヒートマップ $H_{ij}(p, \alpha)$ を

生成する．このとき，DETR の最終層の出力特徴マップを A_{ijk} とし，オブジェクトクエリ p と拡張率 α に対する重みを $w_{ijk}(p,\alpha)$ とする．最後に，式 (13) を用いて，各拡張率 α に対して生成したヒートマップ $H_{ij}(p,\alpha)$ の平均を求め，最終的なヒートマップ $H'_{ij}(p)$ を獲得する．以上の手順により，DETR に対して提案手法を適用することができる．これにより，DETR が物体を検出する際の判断根拠をより解釈しやすい形で可視化することを見込む．

5. 評価実験

本実験では，従来の ODAM と提案手法の可視化結果を比較し，提案手法の有効性を検証する．実験では物体検出モデルとして DETR を採用し，クラス信頼度が 0.7 以上の物体に対して ODAM による可視化を行う．Backbone には ResNet-50 を使用する．Bounding Box の拡張率 α は 1 から 3 倍まで 0.25 倍刻みとして拡張する．

5.1 定量的評価

本章では提案手法の定量的評価として，Energy-based pointing game と Object Discrimination Index を用いて解釈容易性を評価する．

5.1.1 Energy-based pointing game

Energy-based pointing game (EBPG) [12] は判断根拠の可視化結果の解釈容易性を評価する際に使用される．EBPG は，顕著性マップの最大点を抽出し，その最大点が物体の Bounding Box 内に存在するかどうかを確認する従来の pointing game の概念をエネルギーベースの視点で扱うものである．この手法では，最大点のみを使用するのではなく，顕著性マップのエネルギーが検出物体の Bounding Box 内にどれだけ集中しているか評価する．具体的には，まず検出物体の Bounding Box で入力画像を 2 値化し，内側の領域を 1，外側の領域を 0 に割り当てる．次に，生成された顕著性マップの各ピクセルと Bounding Box 内のピクセルごとに乗算し，検出物体の Bounding Box 内に Attention がどれだけあるかを合計する．EBPG は式 (14) のように定義される．

$$\mathrm{EBPG} = \frac{\sum_{(i,j)\in \mathrm{bbox}} L_c(i,j)}{\sum_{(i,j)\in \mathrm{bbox}} L_c(i,j) + \sum_{(i,j)\notin \mathrm{bbox}} L_c(i,j)} \tag{14}$$

ここで $L_c(i,j)$ はクラス c に対する位置 (i,j) での顕著性マップの値を表し，bbox は検出物体の Bounding Box 内のピクセルの集合を表す．分子は Bounding Box 内の顕著性マップのエネルギー合計を，分母は顕著性マップ全体のエネルギー合計を表し，これによりエネルギーの割合が正規化される．

5.1.2 Object Discrimination Index

Object Discrimination Index（ODI）[9] は，ヒートマップが生成するエネルギーが，検出物体以外の物体にどの程度漏れ出しているかを測定する指標である．この指標は，特定の検出物体以外の全ての物体の Bounding Box 内の Attention の割合として，画像内の全物体に対する Attention の総量に対して定義される．背景領域はこの計算から除外される．ODI は式 (15) のように定義される．

$$\mathrm{ODI} = \frac{\sum_{(i,j)\in \text{other objects}} H(i,j)}{\sum_{(i,j)\in \text{all objects}} H(i,j)} \tag{15}$$

ここで，$H(i,j)$ はピクセル位置 (i,j) におけるヒートマップのエネルギーを表し，「other objects」は検出物体以外の全ての物体の Bounding Box 内の領域を指し，「all objects」は画像内の全物体の Bounding Box 内の領域を指す．これにより，ODI は検出物体に関連する Attention がどの程度他の物体に漏れているかを表す割合として計算される．

5.2 EBPG を用いた評価結果

表 1 に異なる拡張率 α を組み合わせた場合の可視化について，EBPG を用いた定量的評価の結果を示す．本実験では，拡張率 α の最適な組み合わせを全探索により求めた．全探索の結果より，Bounding Box の拡張率 α(2.5, 2.75, 3.0) としたとき，提案手法は，従来手法と比較して最大で約 5.16*pt* の精度向上を確認した．つまり，物体の正確な位置と重要な特徴をより効果的に捉え，モデルの判断根拠の解釈容易性が高くなったといえる．

表 1: EBPG を用いた定量的評価

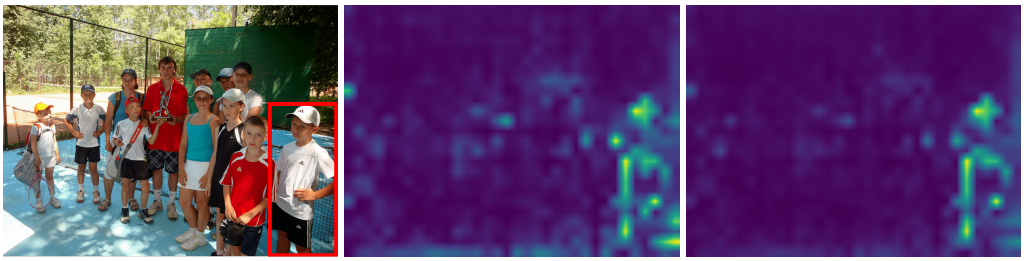
Method	EBPG (%)
従来手法	33.12
提案手法：α(2.5, 2.75, 3.0)	**38.28**

5.3 ODI を用いた評価結果

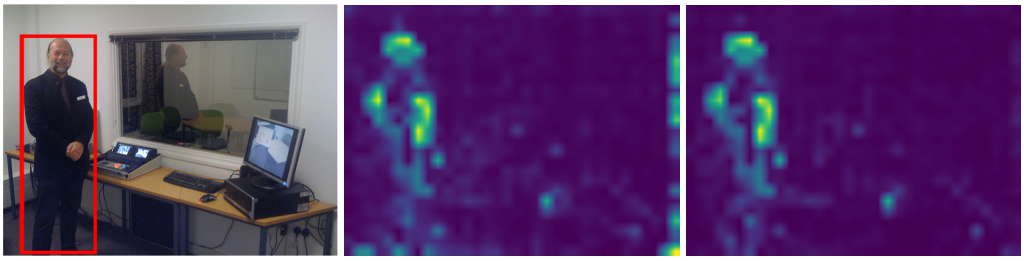
表 2 に異なる拡張率 α を組み合わせた場合の可視化について，ODI を用いた定量的評価の結果を示す．本実験では，拡張率 α の最適な組み合わせを全探索により求めた．全探索の結果より，Bounding Box の拡張率 α(2.5, 2.75, 3.0) としたとき，提案手法は，従来手法と比較して最大で約 5.16*pt* の精度向上を確認した．提案手法は，従来手法と比較して最大で約 7.62pt の精度向上を確認した．つまり，ヒートマップのエネルギーが対象物体内に集中しており，他の物体や背景へのエネルギーの漏れが少ないことから，生成されたヒートマップが対象物体に対して高い識別能力を持っているといえる．

5.4 定性的評価

図 6 に従来手法と提案手法の可視化結果を示す．提案手法の Bounding Box の拡張率 α は 2.5，2.75，3.0 倍とし



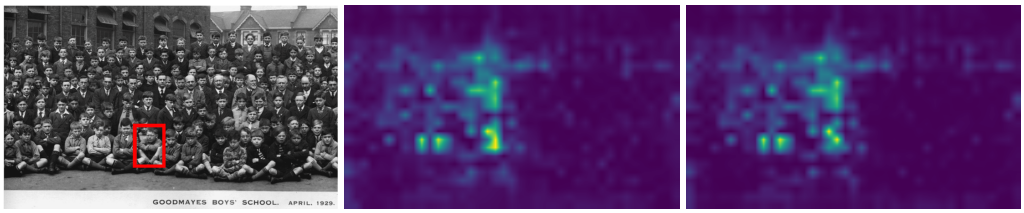
(a) 物体検出結果　(b) 従来手法　(c) 提案手法



(d) 物体検出結果　(e) 従来手法　(f) 提案手法

図 6: 提案手法により精度向上が確認された結果

(a) 物体検出結果　(b) 従来手法　(c) 提案手法



(d) 物体検出結果　(e) 従来手法　(f) 提案手法

図 7: 提案手法により精度向上が確認されなかった結果

表 2: ODI を用いた定量的評価

Method	ODI (%)
従来手法	55.59
提案手法：α(2.5, 2.75, 3.0)	**47.97**

た際の結果である．図 6(b)，図 6(e) に示される通り，従来の ODAM 手法では検出物体以外の箇所を誤って強調し，検出物体以外の箇所や，図 6(b) では画像下部，図 6(e) では画像の右端に帯状のノイズが見られる．これに対し，図 6(c)，図 6(f) に示す提案手法の可視化結果は，検出物体に注視した結果となっており，帯状のノイズの減少も確認された．結果として，提案手法は従来手法に比べて注視物体以外への着目を著しく減少させることができた．一方で，図 7(a) に示すような，画像全体に占める物体の大きさが極端に大きいときには，図 6(c) に示すように，可視化結果の変化は見られなかった．また，図 7(d) に示すような集合写真のように多数の物体を含む画像のうち，特に小物体に対する可視化結果は，図 7(f) に示すように，従来手法と比較してあまり変化は見られなかった．

6. おわりに

本稿では，解釈容易性の向上を目的として ODAM の改良手法を提案した．評価実験では，従来手法と比較して解釈容易性が向上したことを確認した．今後は，検出物体の大きさに応じてより詳細な評価を行い，得られた複数の可視化結果の統合方法を検討する．また，忠実度の維持のた

めに，勾配計算の手法の変更を検討する．

参考文献

[1] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” In Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.

[2] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580-587, 2014.

[3] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788, 2016.

[4] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C.-Y. Fu, and A. C. Berg, “SSD: Single Shot MultiBox Detector,” In European Conference on Computer Vision, pp. 21-37, 2016.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention Is All You Need,” In Advances in neural information processing systems, vol. 30, pp. 5998–6008, 2017.

[6] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-End Object Detection with Transformers,” In Computer Vision - ECCV, pp. 213–229, 2020.

[7] S. Jain and B. C. Wallace, “Attention is not Explanation,” In Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 2019.

[8] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” In Proceedings of the IEEE International Conference on Computer Vision, pp. 618–626, 2017.

[9] C. Zhao and A. B. Chan, “ODAM: Gradient-based Instance-Specific Visual Explanations for Object Detection,” In ICLR, 2023.

[10] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” In Proceedings of the IEEE, vol. 106, no. 12, pp. 2509–2518, 2016.

[11] T.-Y. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, “Microsoft COCO: Common Objects in Context,” In European Conference on Computer Vision (ECCV), pp. 740–755, 2014.

[12] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu, “Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks,” In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 24–25, 2020.