IG-ODAM: Integrated Gradients による 物体検出モデルの判断根拠の可視化

仲井 悠真† 平川 翼† 山下 隆義† 藤吉 弘亘†
†中部大学
E-mail: yuma@mprg.cs.chubu.ac.jp

1 はじめに

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など幅広い分野で広く用いられている．特に Transformer[1] 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．しかし，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．この課題に対応するため，説明可能な AI（XAI: Explainable AI）の分野では，モデルの判断根拠を視覚的に可視化する多様な手法が提案されている．これらの手法は，大きく摂動ベースと勾配ベースの二種類に分類される．摂動ベースの手法は，入力画像を部分的に変化させ，その変化がモデルの出力に及ぼす影響を観察することで判断根拠を推測する．代表的な手法として，Local Interpretable Model-agnostic Explanations (LIME) [2] や D-RISE[3] などが挙げられる．しかし，これらは摂動生成時のランダム性や計算コストの高さ，摂動による入力画像の改変が本来の特徴を破壊する可能性があるという課題を抱えている．勾配ベースの手法は，モデルの出力に対する入力の勾配を利用して重要な特徴領域を可視化する．代表的な手法である Gradient-weighted Class Activation Mapping（Grad-CAM）[4] やその派生手法 [5, 6, 7] は主に画像分類タスクを対象としており，物体検出タスクにおけるインスタンスごとの説明には適していない．物体検出では，検出された各物体インスタンスに対する個別の判断根拠の可視化が求められる．そこで，物体検出タスクに特化した勾配ベースの手法として，Object Detector Activation Maps (ODAM) [8] が提案された．ODAM は物体検出器の出力に対する特徴マップの勾配を用いて，各物体が重要とする領域をヒートマップとして可視化する．しかし，勾配ベースの可視化手法は，感度の公理や実装不変性の公理 [9] を完全には満たさないため，局所的な勾配変動の影響を受け，重要な特徴が過小評価される可能性がある．感度の公理とは，重要な特徴にわずかな変化が生じた際にモデル出力がそれに対応して変化することを要求するものであり，実装不変性の公理とは，モデルが数学的に等価な変換（実装上の違い）があっても説明が一貫していることを保証するものである．これらの公理を満たす手法として，Integrated Gradients[9] が提案されている．

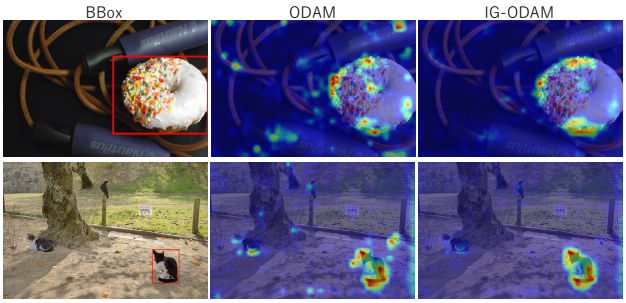


図 1: ODAM と IG-ODAM の比較．提案手法である IG-ODAM は，より解釈性が高い視覚的説明性を持つ．

本論文では，この課題を解決するため，Integrated Gradients を ODAM に組み込んだ IG-ODAM を提案する．従来，Integrated Gradients は画像全体に対する寄与を評価する手法として知られている．しかし，物体検出タスクでは，インスタンスとなる各物体の特徴を個別に可視化する必要があるため，その適用は困難とされている．これは，Integrated Gradients がクラスごとの寄与を評価する手法であり，バウンディングボックス単位での局所的な特徴抽出が難しいことに起因する．提案手法では，類似度を用いたインスタンス間マッチングを導入することで，各物体に対応する局所的な特徴抽出を可能とした．これにより，提案手法である IG-ODAM は，ODAM の物体ごとの説明能力と Integrated Gradients が持つ高い信頼性を兼ね備え，感度の公理と実装不変性の公理を満たしながら，より正確で信頼性のある視覚的説明を実現する．

2 関連研究

本章では関連研究として，勾配ベースの判断根拠の可視化手法と ODAM についてまとめる．

2.1 勾配ベースの可視化手法

深層ニューラルネットワーク（DNN）の判断根拠を可視化する手法の一つに，逆伝播を用いた勾配ベース

の手法がある．この手法では，モデルの出力に対する入力特徴や中間層の活性値の勾配を計算することで，モデルの判断根拠を特定する．Simonyan ら [10] は，モデルの信頼度スコアと入力特徴の間の勾配を計算し，予測に強く影響を与える領域を特定する手法を提案した．その後，解釈の安定性を向上させるため，逆伝播時に負の勾配を抑える Guided Backpropagation[11] などの改良手法が提案された．

この分野における重要な進展として，Integrated Gradients[9] の提案が挙げられる．Integrated Gradients は，感度の公理と実装不変性の公理を満たすアプローチを導入することで，逆伝播による可視化手法の理論的基盤を確立している．

2.2 ODAM

ODAM は物体検出モデルの予測根拠をインスタンスごとに可視化する勾配ベースの手法である．クラススコアやバウンディングボックスの座標といった予測要素の重要度をヒートマップとして可視化し，モデルが各物体をどのように認識したかを示すことができる．

物体検出モデルは入力画像 I に対してインスタンスごとに複数の予測を生成する．各予測 p は，クラススコア $s(c_p)$ とバウンディングボックス $B(p) = (x_1^{(p)}, y_1^{(p)}, x_2^{(p)}, y_2^{(p)})$ から構成される．特定のインスタンス p について，クラススコアやバウンディングボックスの座標などの予測される物体属性のスカラー値 $Y^{(p)}$ は，式 (1) に示すように特徴マップの活性値の重み付き和として表現される．

$$Y^{(p)} = \sum_k \sum_{ij} w_{ijk}^{(p)} A_{ijk} \tag{1}$$

ここで，A_{ijk} は特徴マップの k チャネル目における空間位置 (i, j) での活性値を表し，$w_{ijk}^{(p)}$ は A_{ijk} が予測 $Y^{(p)}$ にどの程度寄与するかを示す重要度の重みである．これらの重みは $Y^{(p)}$ の特徴マップに対する勾配から導出され，インスタンスごとの視覚的説明を可能にする．予測に最も影響を与える領域と特徴を可視化するため，ODAM はインスタンスごとのヒートマップ $H_{ij}^{(p)}$ を生成する．このヒートマップは式 (2) に示すように，空間位置 (i, j) における全チャネルからの寄与を集約したものである．

$$H_{ij}^{(p)} = \sum_k w_{ijk}^{(p)} A_{ijk} \tag{2}$$

最終的に，スカラー出力 $Y^{(p)}$ に対応するインスタンス固有のヒートマップ $H^{(p)}$ は，ピクセル重み付けメカニズムを用いて得られる．重要度の重みマップ $w_k^{(p)}$ は，式 (3) に示すように，勾配マップ $\partial Y^{(p)} / \partial A_k$ に局所平滑化演算 Φ を適用することで定義される．

$$w_k^{(p)} = \Phi\left(\frac{\partial Y^{(p)}}{\partial A_k}\right) \tag{3}$$

この過程で，局所平滑化演算 Φ は勾配マップに適用され，ノイズとして解釈される微小な変動や不規則性を抑制しつつ，本質的な空間的・構造的情報を保持する．ODAM は平滑化演算 Φ としてガウシアンフィルタを採用し，カーネルサイズは特徴マップ内の予測物体のサイズに応じて適応的に決定する．平滑化された重みマップ $w_k^{(p)}$ は，対応する特徴マップ A_k と要素ごとに乗算され，最終的なヒートマップ $H^{(p)}$ の構築に寄与する．このヒートマップは，ピクセル重み付けメカニズムを用いて構築され，全特徴マップチャネルからの寄与が集約される．式 (4) に示すように，予測に対する正の影響のみを強調するため ReLU 関数を適用している．

$$H^{(p)} = \mathrm{ReLU}\left(\sum_k w_k^{(p)} \circ A_k\right) \tag{4}$$

ここで，$\circ$ は要素ごとの乗算を表す．

3 提案手法

提案手法である IG-ODAM は，ODAM に Integrated Gradients を組み込むことで，物体検出モデルの解釈性を向上させ，より解釈性の高い視覚的説明を実現する手法である．本手法は，物体検出における予測時にモデルがどの画像領域に着目しているかを視覚的に説明する．さらに本手法では，類似度に基づくインスタンスマッチングを導入することで，積分過程において生成される各補間画像における検出結果と，元画像での検出結果との対応関係を確立する．これにより，物体単位での判断根拠を高い信頼性で可視化できる．図 1 に示すように，本手法は従来の ODAM と比較して，局所的な勾配の変動に対する感度が低減されており，より一貫性のある視覚的説明が得られる．

3.1 Integrated Gradients

逆伝播に基づく特徴帰属手法の主な課題の一つは，特徴帰属に関する理論的保証が欠如していることである．特徴帰属とは，モデルの予測に対する各入力特徴の寄与を定量的に評価することであり，一貫性や理論的な正当性が求められる．ここで，各入力特徴の重要度を示す数値を帰属値と呼ぶ．この課題を解決するため，Integrated Gradients は以下の 2 つの基本的な公理に基づくアプローチとして提案されている．

- 感度の公理：
  入力とベースラインの間で，モデルの予測に影響を及ぼす特定の入力要素が異なる場合，帰属手法はその特徴に非ゼロの帰属値を割り当てなければならない．一方，ニューラルネットワークがある変数に依存しない場合，その帰属値はゼロでなければならず，これにより依存関係の正確な測定が保証される．

- 実装不変性の公理：
  2 つのネットワークが全ての入力に対して同一の出力を生成する場合，帰属手法は同一の帰属値を生成しなければならない．これにより，機能的に等価なモデル間での一貫性が保証される．

Integrated Gradients における重要な要素の一つがベースラインである．ベースラインとは，入力の参照点となる中立的なデータであり，多くの場合，情報を含まない黒画像が使用される．例えば画像認識タスクにおいて，黒画像をベースラインとした場合，各画素の重要度は「黒色から実際の画素値に変化したときに予測出力へ与える影響」として定義される．ベースラインは黒画像以外にも，ぼかし画像やランダムノイズ画像など，タスクに応じて使用可能である．得られた帰属値は，ヒートマップとして可視化される．高い解釈可能性を有することから，Integrated Gradients は，画像認識や自然言語処理を含む様々なタスクで広く採用されている．

3.2 類似度を用いたインスタンスマッチング

画像分類タスクでは，Integrated Gradients などの勾配ベースの説明手法により，画像全体に対する勾配を計算することで各クラスへの特徴の寄与を可視化できる．しかし，物体検出タスクでは，一つの画像内に複数の物体が存在し，それぞれに対して個別の予測（バウンディングボックスとクラススコア）が行われるため，画像全体ではなく検出結果ごとに勾配を計算する必要がある．そこで，物体検出のための Shapley ベースの説明手法である BSED[12] で用いられている類似度指標を適用し，式 (5) に示すように元画像と補間画像間の検出結果の対応関係を確立する．

$$\mathrm{Sim}(d_t, d_{j,k}) = s_{\mathrm{loc}}(d_t, d_{j,k}) \cdot s_{\mathrm{cls}}(d_t, d_{j,k}) \tag{5}$$

ここで，s_{loc} はバウンディングボックスの IoU に基づく位置の類似度を，s_{cls} はクラススコアの類似度を表す．補間画像 X_k における最適な検出結果は，式 (6) に示すように類似度が最大となる検出結果として選択される．

$$f(X_k) = \max_{d_{j,k} \in \phi(X_k)} \mathrm{Sim}(d_t, d_{j,k}) \tag{6}$$

ここで，$\phi(X_k)$ は補間画像 X_k におけるすべての検出結果の集合を表す．この類似度に基づく対応付けにより，Integrated Gradients の計算過程で生成される各補間画像に対して，元画像の検出結果に対応する適切な検出結果を特定することが可能となる．この手法により，ODAM のインスタンス固有の可視化能力を維持しつつ，Integrated Gradients の理論的利点を活かした，より正確で信頼性の高い特徴帰属を実現できる．

3.3 IG-ODAM

本節では，提案手法 IG-ODAM を用いた物体検出モデルの判断根拠の可視化手法について述べる．

図 2 に提案手法の概要を示す．本手法では，ベースライン画像（黒画像）から入力画像までの直線経路に沿って補間画像を生成し，各補間画像における勾配情報を積分することで，より安定かつ信頼性の高い視覚的説明を可能にする．

物体検出モデルは，入力画像 I から物体の集合 p を予測する．クラススコア $s(C_p)$ およびバウンディングボックス $B(p) = (x_1^{(p)}, y_1^{(p)}, x_2^{(p)}, y_2^{(p)})$ で表現される．モデルの出力スコア $Y^{(p)}$ は，特徴マップの線形結合として式 (7) で示される．ここで，特徴マップの i, j は空間的位置を，k は特徴チャネルを示す．

$$Y^{(p)} = \sum_k \sum_{i,j} w_{ijk}^{(p)} A_{ijk} \tag{7}$$

ここで，$w_{ijk}^{(p)}$ は特徴マップ A_{ijk} に対する重要度の重みを表す．従来の ODAM では局所的な勾配から重みを計算するが，Integrated Gradients を用いる本手法では，ベースライン画像 I' から入力画像 I までの経路に沿った勾配を積分することで特徴の重要度を評価する．中間画像 I_α は式 (8) で定義される．

$$I_\alpha = I' + \alpha(I - I') \tag{8}$$

ここで，I' はベースライン画像であり，$\alpha \in [0, 1]$ は I' と I の間の補間を決定するパラメータである．Integrated Gradients は式 (9) によって定義されるが，この積分を解析的に計算することは一般に困難である．そこで，式 (10) に示すリーマン和による離散近似を用いる．

$$IG(I) = (I - I') \int_0^1 \frac{\partial F(I_\alpha)}{\partial I} d\alpha \tag{9}$$

$$IG(I) \approx (I - I') \sum_{m=1}^{M} \frac{\partial F(I_{\frac{m}{M}})}{\partial I} \times \frac{1}{M} \tag{10}$$

ここで，M は積分近似のためのステップ数であり，精度と計算コストのバランスを考慮して $M = 50$ とした．ステップ数を増やすと精度は向上するが，計算コストが線形に増加する．離散近似を用いて，重要度の重み $w_k^{(p)}$ は式 (11) で定義され，その離散近似は式 (12) で与えられる．

$$w_k^{(p)} = \Phi\left(\int_0^1 \frac{\partial Y^{(p)}}{\partial A_k(I_\alpha)} d\alpha \right) \tag{11}$$

$$w_k^{(p)} \approx \Phi\left(\sum_{m=1}^{M} \frac{\partial Y^{(p)}}{\partial A_k(I_{\frac{m}{M}})} \times \frac{1}{M} \right) \tag{12}$$

これらの重要度の重みを用いて，物体 p に対するヒートマップ $H^{(p)}$ は式 (13) で定義される．

$$H_{ij}^{(p)} = \sum_k w_k^{(p)} A_{ijk} \tag{13}$$

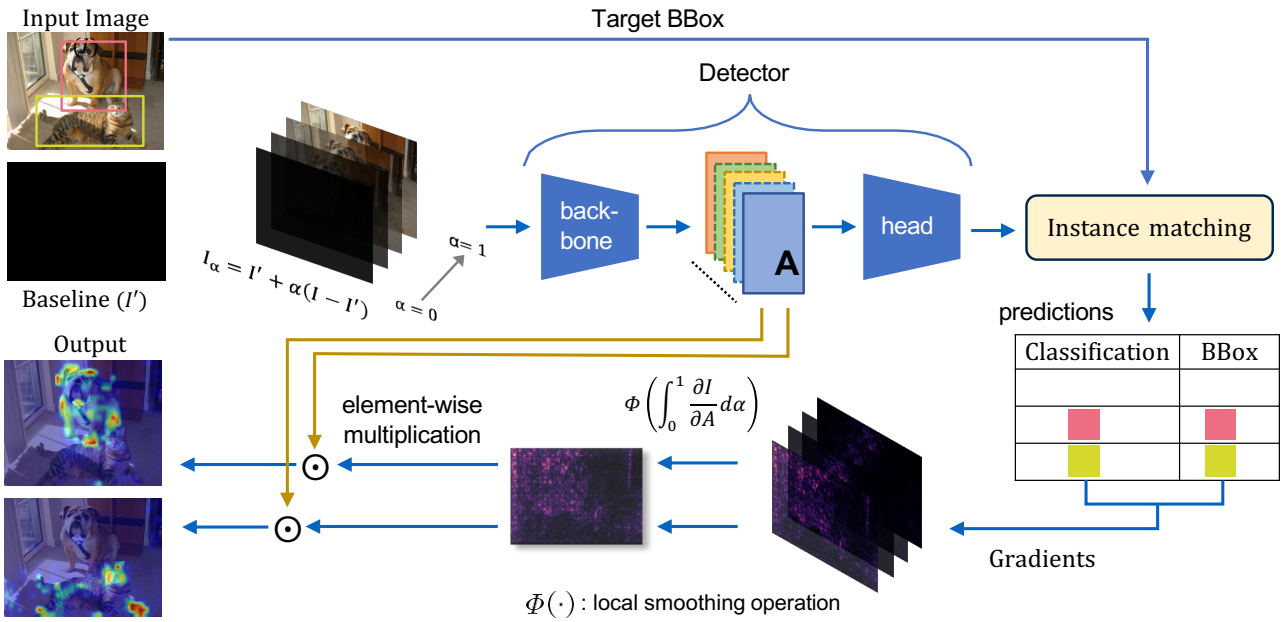


図 2: 提案手法の概略図．ベースライン画像を黒画像とし，ベースラインと入力画像間を直線経路で補間した場合の処理を示す．

さらに，負の寄与を除去するため ReLU 活性化関数を適用すると，最終的なヒートマップは式 (14) で得られる．

$$H^{(p)} = \mathrm{ReLU}\left(\sum_k w_k^{(p)} \circ A_k\right) \tag{14}$$

ここで，$\circ$ は要素ごとの乗算を表す．

4 評価実験

提案手法の解釈性と忠実性を定量的に評価するため，比較実験を行う．Integrated Gradients のベースライン画像には，すべての画素値を 0 に設定した黒画像を使用し，50 枚の中間画像を用いる．評価には，CNN[13] と Transformer に基づく物体検出モデルを採用した．CNN ベースのモデルとして Fully Convolutional One-Stage Object Detection（FCOS）[14] を，Transformer ベースのモデルとして End-to-End Object Detection with Transformers（DETR）[15] を用いた．両モデルはバックボーンに ResNet-50[16] を使用し，FCOS では Feature Pyramid Network（FPN）[17] を組み込んだ．評価には MS COCO データセット [18] を使用し，Grad-CAM，Grad-CAM++，D-RISE[19]，ODAM との定性的な比較を行った．また，ユーザ評価を実施し，提案手法の説明性向上を客観的に評価した．実験はすべて PyTorch フレームワーク [20] で実装し，NVIDIA RTX A6000 GPU 上で実施した．

4.1 判断根拠の忠実性評価

Deletion，Insertion[21] は，Grad-CAM や RISE などの可視化手法が，ニューラルネットワークの顕著性マップにおいて重要な領域をどの程度正確に特定できているかを評価する指標である．Deletion では，顕著性マップが示す重要度の高い順に，入力画像の画素をランダムな値に置換していく．この過程でモデルの予測信頼度が急速に低下する場合，置換された画素がモデルの判断において実際に重要な役割を果たしていたことを意味する．一方，Insertion では，最初に黒画像やノイズ画像など，内容を持たない画像を用意し，そこに顕著性マップで示された重要度の高い領域から順に，元の入力画像の情報を徐々に追加していく．この過程でモデルの予測信頼度が急速に上昇すれば，追加された領域が実際にモデルの判断に重要であることを意味し，顕著性マップが適切に重要な部分を特定できていると判断できる．

本実験では，Deletion，Insertion 評価を行う際，ベースライン画像から入力画像へ向かう線形補間により合計 100 枚の補間画像を作成し，各ステップで画素を削除または追加してモデルの予測変化を測定した．また，検出物体の大きさに応じて適切な評価を行うため，各ステップで削除または追加する画素数は，対象のバウンディングボックスの面積に基づいて動的に決定した．具体的には，バウンディングボックスの面積を 100 分割し，その値を 1 ステップあたりに操作する画素数として設定している．すべての物体に対して得られたスコアの平均値を，各手法の最終的なスコアとして用いた．

表 1 に各可視化手法の Deletion および Insertion の結果を示す．Deletion 値は低いほど，Insertion 値は高いほど，モデルの判断根拠を忠実に反映していることを意味する．Grad-CAM および Grad-CAM++は，Deletion

表 1: COCO データセットにおける定量的分析の結果．最良の値を太字で示す：Deletion は値が低いほど，Insertion は値が高いほど良い．

Method	Del.↓	Ins.↑
Grad-CAM	92.79	36.78
Grad-CAM++	92.52	36.18
D-RISE	73.35	43.35
ODAM	72.68	50.33
IG-ODAM	**70.4**	**57.19**

値が約 92 と高く，Insertion 値が約 36 と低いため，モデルの注視領域を十分には反映できていないことがわかる．一方，D-RISE は Deletion が 73.35，Insertion が 43.35 と，Grad-CAM 系の手法より良好な結果を示した．さらに，ODAM は Deletion が 72.68，Insertion が 50.33 と改善されている．提案手法 IG-ODAM は，Deletion が 70.4，Insertion が 57.19 と，すべての指標で既存手法を上回る結果を示した．結果より，提案手法は従来手法と比較して，モデルの判断根拠をより正確に捉えていることが明らかである．

4.2 判断根拠の解釈性評価

提案手法である IG-ODAM の解釈性を，Grad-CAM，Grad-CAM++，D-RISE，およびベースライン手法 ODAM と比較した．図 3 に示す可視化結果から，IG-ODAM は ODAM と比較してノイズが大幅に低減され，検出された物体に対するより明確な視覚的説明が得られた．ベースライン手法の ODAM は検出物体を適切に捉えていたが，IG-ODAM では物体の重要な特徴をさらに強調できたことが確認された．

既存手法との比較において，Grad-CAM と Grad-CAM++は活性化領域が広範囲に及び，D-RISE は高いノイズレベルと不安定な説明を生成する傾向が見られた．一方，IG-ODAM は検出物体の境界を明確に示しながら，安定した説明を示した．特に複数のインスタンスが存在するシーンでは，IG-ODAM は各検出物体に対して明確に分離された説明を生成し，同一クラス内での説明の一貫性も向上した．

図 4 に，Transformer ベースの DETR と CNN ベースの FCOS に対する IG-ODAM と ODAM の可視化結果を示す．図 4(a)，図 4(b) は DETR の物体検出結果を，図 4(c)，図 4(d) は FCOS の物体検出結果をそれぞれ可視化したものである．これらの結果から，両モデルの推論特性の違いが明確にあることが分かる．DETR は物体の輪郭や離れた部分の情報を統合して推論を行う傾向がある一方，FCOS は局所的な特徴を重視する傾向が確認された．さらに，提案手法である IG-ODAM は，従来の ODAM と比較してこれらの特徴をより明瞭に可視化できた．具体的には，DETR に対する結果（図 4(a)，4(b)）では広域的な特徴の統合過程が，FCOS に対する結果（図 4(c)，4(d)）では局所的特徴の強調がより顕著に表現されている．以上の結果は，IG-ODAM が Transformer ベース，CNN ベースの両アーキテクチャに対して適用可能であり，かつそれぞれのモデルの特徴的な推論過程を ODAM よりも効果的に可視化できることを示している．

4.3 ユーザーの信頼性評価

提案手法 IG-ODAM が生成する視覚的説明の解釈性を評価するため，ユーザー評価実験を実施した．本実験では，MS COCO データセットの検証セットから 30 クラスを選択し，正しく検出された物体について評価を行った．比較対象として，ODAM，Grad-CAM，Grad-CAM++，D-RISE を用いた．

実験では，各物体について 5 つの手法で生成されたヒートマップを被験者に提示し，物体検出過程の説明としての適切さに基づいてランク付けを依頼した．41 名の被験者から，合計 630 件の有効回答を得た．

評価結果を表 2 に示す．分析の結果，IG-ODAM は全試行において平均ランク 1.57 と最も高い評価を受け，従来手法を上回る性能を示した．これらの結果は，IG-ODAM が従来手法と比較して，より高い視覚的説明性を持ち，物体検出過程に対するユーザーの信頼性を高めることを示唆している．

結果の統計的有意性を確認するため，カイ二乗検定およびウィルコクソンの符号順位検定を実施した．カイ二乗検定により，5 つの手法間のランク分布に有意な差異が確認された（χ^2，$p < 0.001$）．さらに，ウィルコクソンの符号順位検定により，IG-ODAM が他のすべての手法に対して有意に優れていることが確認された（$p < 0.001$）．これらの結果は，IG-ODAM は解釈可能な視覚的説明を生成し，ユーザーの信頼性向上に効果的であることが統計的に示された．

表 2: ユーザーの信頼性評価に基づく可視化手法の比較：各ランクの獲得割合と平均ランク（AR）

Method	1st	2nd	3rd	4th	5th	AR
IG-ODAM	**417**	126	47	24	16	**1.57**
ODAM	102	414	74	33	7	2.09
Grad-CAM	24	42	255	190	119	3.54
Grad-CAM++	67	34	229	285	15	3.23
D-RISE	20	14	25	98	473	4.57

4.4 IG-ODAM における積分ステップ数

Integrated Gradients におけるステップ数の設定は，これまでデータセットやモデルごとに経験則に基づいて決定されてきた．しかし，このような固定的な設定は，解析結果の信頼性に影響を及ぼす可能性が指摘さ

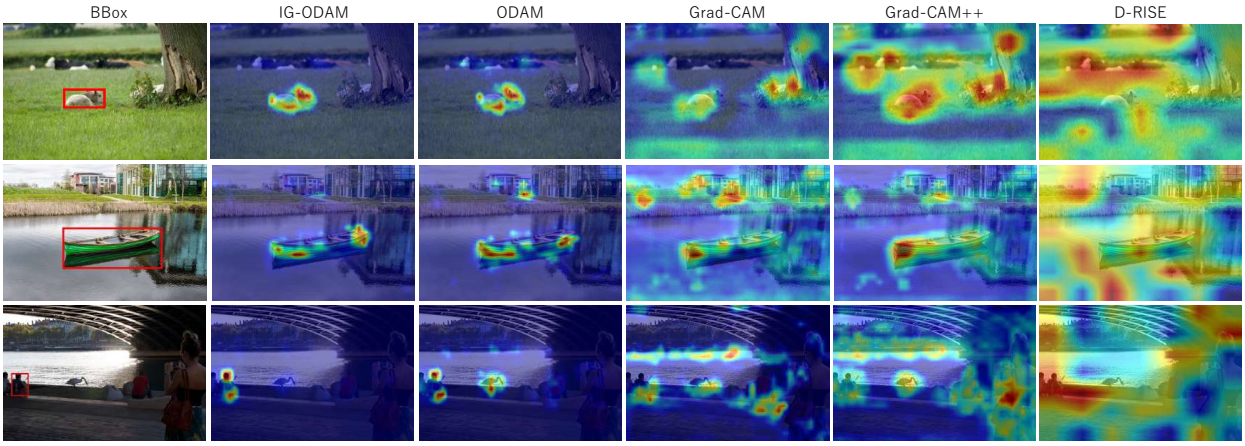
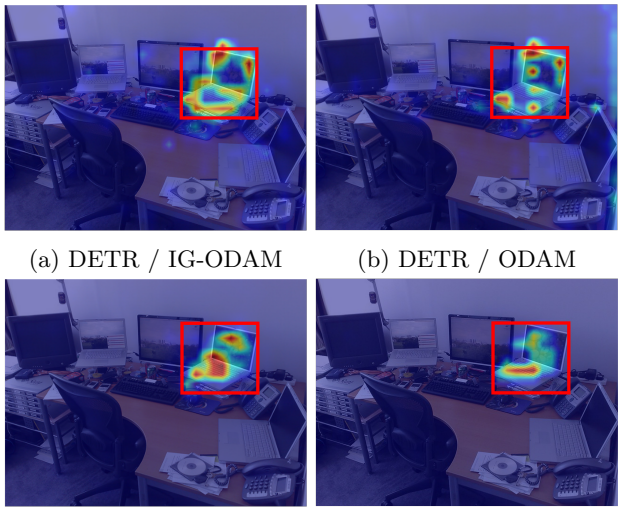


図 3: COCO データセットにおける DETR の物体検出結果の可視化:検出結果を含む入力画像（左端）と，IG-ODAM，ODAM，Grad-CAM，Grad-CAM++，D-RISE による可視化マップ.



(a) DETR / IG-ODAM　(b) DETR / ODAM

(c) FCOS / IG-ODAM　(d) FCOS / ODAM

図 4: DETR および FCOS に IG-ODAM と ODAM を適用した物体検出結果の可視化結果

れている [22]．本研究では，IG-ODAM における最適なステップ数の決定方法を確立するため，補間画像の枚数が可視化の精度と計算コストに与える影響を評価する．

4.4.1　評価手法

従来は補間画像の枚数を 50 枚に固定することが一般的であったが，本実験では 0 枚（従来の ODAM）から 300 枚まで 10 ステップ刻みで変化させ，その影響を評価した．評価指標として，Sundararajan ら [23] の提案に基づき，相対誤差が 5%以内に収まる最小ステップ数を理想的なステップ数と定義した．ODAM マップの相対誤差 $\varepsilon(N)$ を式 (15) のように定義する．

$$\varepsilon(N) = \frac{1}{|\Omega|} \sum_{(i,j)\in\Omega} \frac{|I_N(i,j) - I_{\text{ref}}(i,j)|}{|I_{\text{ref}}(i,j)|} \times 100 \quad (15)$$

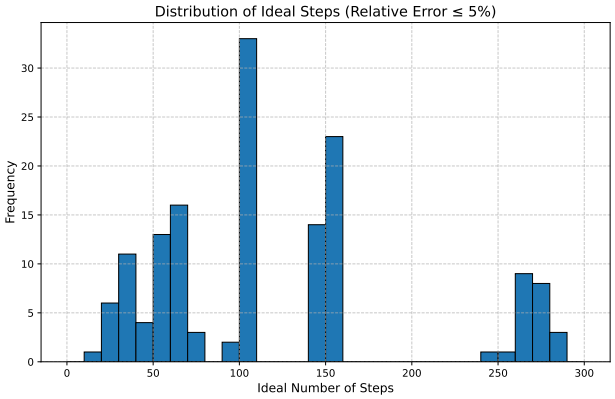


図 5: 相対誤差が 5%以下となる理想的なステップ数の分布．横軸はステップ数，縦軸は頻度を表す．

ここで，$I_N(i,j)$ は補間画像枚数 N に対応する ODAM マップの画素値，$I_{\text{ref}}(i,j)$ は 300 ステップにおける ODAM マップの画素値（基準値），Ω は画像の全画素集合を表す．理想的なステップ数 N^* は，式 (16) のように定義する．

$$N^* = \min\{N \mid \varepsilon(N) \leq 5\} \quad (16)$$

4.4.2　分析結果と考察

図 5 に示す理想的なステップ数の分布から，多くのケースにおいて 50〜100 ステップの範囲で相対誤差が 5%以下に収束することが確認された．特に 100 ステップ付近に見られる顕著なピークは，この値が一般的に適切な選択となることを示唆している．一方で，150 ステップ付近にも別のピークが観察され，一部のデータでは 100 ステップでは十分な精度が得られないことが明らかになった．この結果は，Integrated Gradients の積分精度がデータの性質に強く依存することを示している．

また，計算時間はステップ数に対して線形に増加することも確認された．1 インスタンあたりの追加計算時間は約 64.85 秒であり，300 ステップでは従来の 50 ステップと比較して約 5 倍の計算時間を要する．これらの結果より，Integrated Gradients で作成する補間画像の枚数は，データの特性に応じて適応的にステップ数を決定する手法が望ましいと考える．

5 おわりに

本論文では，物体検出モデルの解釈性を向上させるため，ODAM に Integrated Gradients の経路積分計算を組み込んだ IG-ODAM を提案した．実験結果から，IG-ODAM は従来手法と比較して，より焦点を絞った明確なインスタンス固有の判断根拠を可視化可能であることが示された．本研究の主な貢献は以下の通りである：

1. Integrated Gradients の物体検出タスクへの初めての適用
2. 補間画像を効果的に処理する類似度ベースのインスタンスマッチング手法の提案
3. 定量的指標とユーザーの信頼性評価の両面における性能向上の実証

提案手法は，感度の公理や実装不変性の公理といった理論的性質を満たすことで，信頼性の高い XAI モデルの構築に寄与する．一方で，Integrated Gradients の導入により，補間画像ごとの勾配計算が必要となり，計算コストの増加が課題として明らかになった．さらに，積分ステップ数の分析から，最適なステップ数がデータの特性に依存することが示された．この知見を踏まえ，今後は，インスタンスごとの特徴や重要度に応じて積分ステップ数を動的に決定する手法を検証する．

参考文献

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[2] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[3] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. In *British Machine Vision Conference (BMVC)*, pages 1–13, 2018.

[4] R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization, 2016. This version was published in International Journal of Computer Vision (IJCV) in 2019; A previous version of the paper was published at International Conference on Computer Vision (ICCV'17).

[5] Aditya Chattopadhyay, Anirban Sarkar, Prantik Howlader, and Vineeth N. Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. *CoRR*, abs/1710.11063, 2017.

[6] Haofan Wang, Mengnan Du, Fan Yang, and Zijian Zhang. Score-cam: Improved visual explanations via score-weighted class activation mapping. *CoRR*, abs/1910.01279, 2019.

[7] Mohammed Bany Muhammad and Mohammed Yeasin. Eigen-cam: Class activation map using principal components. In *2020 International Joint Conference on Neural Networks (IJCNN)*, page 1–7. IEEE, July 2020.

[8] Chenyang Zhao and Antoni B. Chan. Odam: Gradient-based instance-specific visual explanations for object detection, 2023.

[9] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR, 2017.

[10] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps, 2014.

[11] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net, 2015.

[12] Michihiro Kuroki and Toshihiko Yamasaki. Bsed: Baseline shapley-based explainable detector. *IEEE Access*, 12:57959–57973, 2024.

[13] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, volume 86, pages 2278–2324, 1998.

[14] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection, 2019.

[15] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov,

and Sergey Zagoruyko. End-to-end object detection with transformers, 2020.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[17] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection, 2017.

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015.

[19] Vitali Petsiuk, Rajiv Jain, Varun Manjunatha, Vlad I. Morariu, Ashutosh Mehra, Vicente Ordonez, and Kate Saenko. Black-box explanation of object detectors via saliency maps, 2021.

[20] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019.

[21] Naofumi Hama, Masayoshi Mase, and Art B. Owen. Deletion and insertion tests in regression models, 2023.

[22] Masahiro Makino, Yuya Asazuma, Shota Sasaki, and Jun Suzuki. The impact of integration step on integrated gradients. In Neele Falk, Sara Papi, and Mike Zhang, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 279–289, St. Julian's, Malta, March 2024. Association for Computational Linguistics.

[23] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks, 2017.