Grid-wise-attention を用いた物体検出の視覚的説明性に関する研究

TP20011 木村秋斗　　指導教授：山下隆義

1. はじめに

物体検出は画像上の物体位置とその名称を特定する技術である．自動運転における安全確保には歩行者や車両の検出が欠かせない．特に，レベル 5 の完全自動運転の実現には，検出精度の向上だけでなく，サービスを享受する人が信頼できるように検出結果の判断根拠を示す必要がある．

本研究では，歩行者を検出対象とし，その判断根拠をアテンションマップとして出力する手法を提案する．提案手法は，画像をグリッド状に分割し，各グリッドに対するアテンションマップを取得する．また，提案手法では，取得したアテンションマップを特徴マップに重み付けして位置推定を行うことで，検出精度の向上を実現する．

2. 関連研究

物体検出手法は，物体の位置推定とクラス分類を同時に行う one-stage と，物体の位置推定をした後，各位置に対してクラス分類を行う two-stage がある．one-stage の手法は，高い精度を達成しつつ高速化されており，物体検出の代表的なアプローチとなっている．one-stage 検出手法の 1 つである Center and Scale Prediction(CSP)[1] は，歩行者の中心とスケールを推定して歩行者を検出する．これにより，小さい物体やオクルージョンに頑健な検出が可能である．しかし，CSP を含む従来の物体検出手法は，ネットワークの中間情報を可視化した際に複数の物体を同時に検出するため，どの物体の判断根拠を示したものであるか不明であるという問題がある．

3. 提案手法

本研究では，画像をグリッド状に分割した各グリッドに対するアテンションマップを取得する Grid-wise-attention による物体検出を提案し，アテンションマップから物体検出の判断根拠を示す．

3.1. ネットワーク構造

提案手法のネットワーク構造を図 1 に示す．本手法のネットワーク構造は，Feature extractor，各グリッドに対するクラス確率を求める Classification branch と Grid-wise-attention による重みづけによって各グリッドに対する物体の位置を求める Localization branch から構成される．

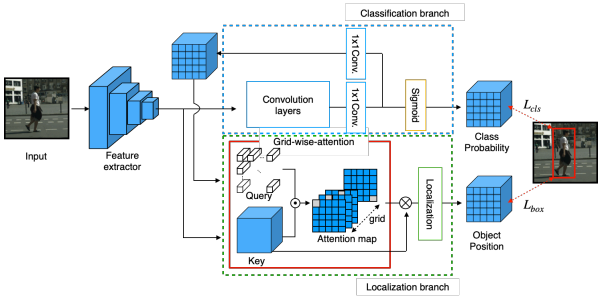

図 1：提案手法のネットワーク構造

はじめに画像をベースネットワークである ResNet-50 に入力し，特徴マップを取得する．次に，特徴マップを $W \times H$ のグリッド状に分割し，Classification branch の畳み込み層とシグモイド関数によりグリッド毎に物体のクラス確率を出力する．このときの特徴マップを Grid-wise-attention 機構において，$W \times H$ 個分の Query，ベースネットワークから取得した特徴マップを Key として，Query と Key の内積によってアテンションマップを $W \times H$ 個分取得する．位置 (i, j) のグリッドに対するアテンションマップ $\mathrm{Attention}_{ij}$ を式 (1) に示す．

$$\mathrm{Attention}_{ij} = \mathrm{softmax}(\boldsymbol{Q}_{ij}\boldsymbol{K}^{\mathrm{T}}) \tag{1}$$

取得される $\mathrm{Attention}_{ij}$ は，Transformer[2] と同様に Query と Key の類似度を表す．この処理を全ての Query について行い，各 Query に対するアテンションマップを取得する．取得したアテンションマップをベースネットワークから取得した特徴マップに重み付けする．これにより，注視領域を考慮した物体検出ができる．そして，Localization branch において，重み付けされた特徴マップから各グリッドに対しての物体位置を出力する．1 つの物体に対して複数の検出結果を出力することを防ぐために，非最大値抑制を用いる．

3.2. 学習方法

提案手法は，出力した各グリッドに対してのクラス確率と物体位置から求めた検出結果に対するクラス確率の損失 L_{cls} と推定位置と正解位置の損失 L_{box} を用いて学習する．クラス分類には Focal loss を使用する．Focal loss は，高い尤度で認識したクラスに対して，小さな重みを損失に乗算する．これにより，認識が困難なクラスを重視するような学習ができる．L_{cls}，L_{box} をそれぞれ式 (2)，式 (3) に示す．

$$L_{cls} = \frac{1}{N^c}\sum_i -(1-\hat{p}_i)^{\gamma}\log(\hat{p}_i) \tag{2}$$

$$L_{box} = \frac{1}{N^r}\sum_i I(g_i = 1) l_r(b_i, t_i) \tag{3}$$

ここで，γ を簡単なクラスに対する重みを調整するパラメータで本研究では $\gamma = 2$ とする．N_k^c を検出結果数，g_i を i 番目の検出結果に対する正解クラス，l_r を smooth L1 Loss，b_i を i 番目の検出結果の物体位置，t_i を i 番目の検出結果に対する正解位置，$I(\cdot)$ を正解と定義されたアンカーのみに制限する指標関数とする．$\hat{p}_i$ は，i 番目の検出結果のクラス確率 p_i を用いて式 (4) の条件で求められる．

$$\hat{p}_i = \begin{cases} p_i & \text{if } g_i = 1 \\ 1 - p_i & \text{otherwise} \end{cases} \tag{4}$$

3.3. 物体検出の注視領域の取得

Grid-wise-attention 機構は位置 (i, j) のグリッドに対するアテンションマップ $\mathrm{Attention}_{ij}$ を取得する．これを特徴マップの全ての位置に対して求め，2 次元に配置することでアテンションマップを求める．本手法の物体検出は，各グリッドに対してのクラス確率と物体位置を基に行うため，グリッド毎に検出結果を出力する．そのため，閾値処理や非最大値抑制を適用した後に物体を含むグリッドに対してのアテンションマップを可視化することで，検出対象毎の判断根拠を取得することができる．

3.4. マルチスケール構造の導入

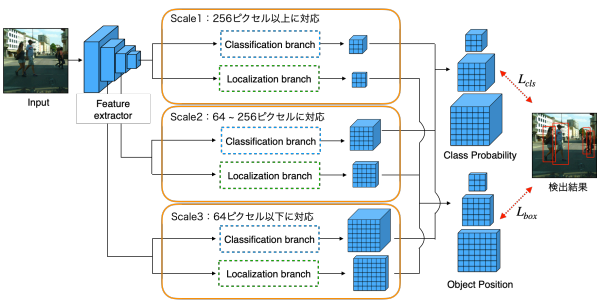

図 2：マルチスケール構造

本手法は，図 2 に示すように異なるスケールのグリッドを用いて多重解像度化する．ベースネットワークから異なるスケールの特徴マップを取得し，それぞれを Classification branch と Localization branch に入力してアテンションマップの取得と物体の検出を行う．これにより，高解像度の特徴マップから小さな物体を検出，低解像度の特徴マップから大きな物体の検出をすることが可能となる．

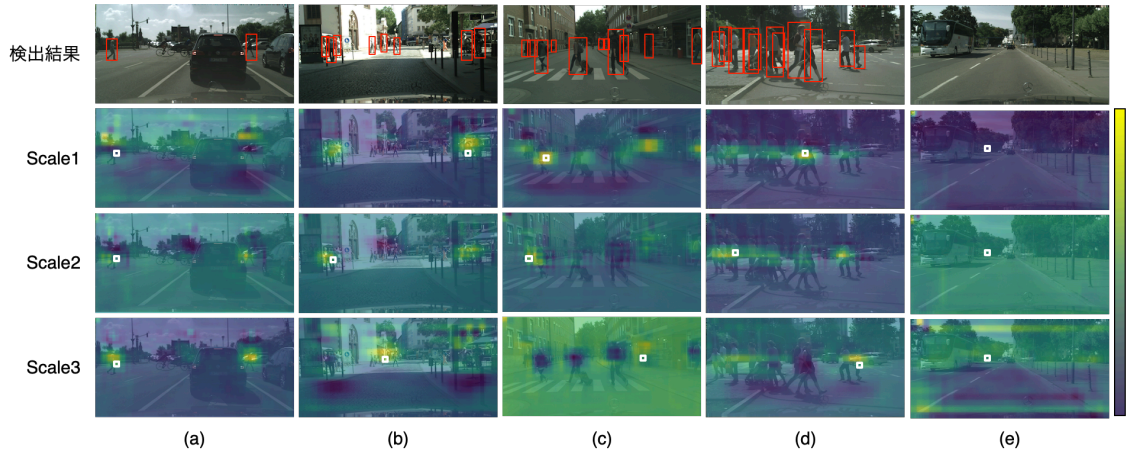

図 4：アテンションマップ可視化結果

4. 評価実験

本実験では，提案手法によるアテンションマップを可視化することで，歩行者検出の判断根拠を示す．また，提案手法の検出精度から，注視領域を考慮した歩行者検出の有効性を調査する．

4.1. 実験概要

学習，評価には，学習用に 2,975 枚，評価用に 500 枚の画像を含む CityPersons データセットを用いる．提案手法は学習回数を 150 エポック，最適化手法を Adam，初期学習率を 1.0×10^{-5} とする．

比較実験では従来の物体検出手法である CSP を用いる．定量的な評価指標には log-average miss rate を用いる．log-average miss rate は 1 画像あたりの誤検出率 (FPPI) を 0.01 から 100 の範囲内から対数スケールで等間隔に取得し，各地点の FPPI に対する未検出率 (Miss Rate) の平均から求める．log-average miss rate が低いほど高精度であることを示す．正解である歩行者の高さが 250 pixel 以上 のものを Large，250 pixel 未満かつ 50 pixel 以上のものを Middle，50 pixel 未満のものを Small する．

また，アテンションマップを可視化し，強く注視する箇所を基に判断根拠を推測する．異なるスケールの特徴マップから取得するアテンションマップを入力特徴マップの解像度が低い順に Scale1，2，3 として，スケール毎のアテンションマップを比較する．

4.2. 検出精度の比較

従来の歩行者検出手法と提案手法の比較結果を表 1 に示す．提案手法は Middle，Small ではそれぞれ CSP に比べ 0.5 ポイント，4.9 ポイント増加しているが，Large では 12.4 ポイント低下しており，全体では 7.1 ポイント低下し，精度が向上している．

表 1：検出精度の比較（log-average miss rate）

	All	Large	Middle	Small
CSP[1]	20.2	23.2	**5.8**	**15.1**
提案手法	**13.1**	**10.8**	6.3	20.0

また，DET カーブを図 3 に示す．提案手法は，CSP より DET カーブが原点に近いため，高性能であるといえる．

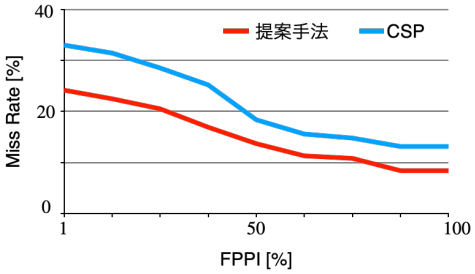

図 3：DET カーブ

検出した歩行者のアテンションマップを図 4 に示す．図のアテンションマップは白枠のグリッドに対してのアテンションマップであり，青色に近いほど弱く，黄色に近いほど強く反応していることを示す． 図 4(a) から (d) のアテンションマップから，歩行者と推測される対象の体付近を強く注視していることがわかる．また，図 4(a)，(b) では，各スケール毎に対応する大きさの歩行者に対して強く注視しており，図 4(c)，(d) の Scale3 では，近辺の歩行者に対して背景よりも弱く注視している．このことから，本手法は対応しない大きさ歩行者の誤検出を防ぎつつ，歩行者の体に注目して検出していることがわかる．一方で，図 4(d) の Scale1，2 のアテンションマップは，同じ歩行者に対して強く注視しているため，検出結果が重複してしまっている．Scale が異なることで，非最大値抑制で削除しきれず誤検出となった．そのため，提案手法が CSP より Middle において低い精度となった．また，図 4(e) の遠方の小さいサイズの歩行者に対するアテンションマップは，小さな物体の検出ができる Scale3 においても歩行者に対して注視できていない．このことから，提案手法が CSP より Small において精度が低いのは，一部の小さな物体に対してアテンションが低いためであると考える．

5. おわりに

本研究では，グリッド状に分割した各地点に対するアテンションマップを取得する Grid-wise-attention による歩行者検出手法を提案した．従来の歩行者検出手法との比較実験では，提案手法が全体で 7.1 ポイント精度が向上したことから，注視領域を考慮した重み付けによって高精度な歩行者検出ができることを示した．また，注視領域から歩行者検出の判断根拠を示した．グリッドに対するアテンションマップから本手法による歩行者検出は，対応する大きさの歩行者の体が判断根拠となることがわかり，誤検出や未検出が発生した要因を示すことができた．今後は，各グリッドに位置の情報を与えることでより対象物体を注視するアテンションマップの獲得と複数クラスを検出する際の判断根拠の可視化を行う．

参考文献

[1] W. Liu, *et al.*, “Center and Scale Prediction: A Box-free Approach for Pedestrian and Face Detection”,CVPR, 2019.

[2] A. Vaswani, *et al.*, “Attention is all you need”,NIPS, 2017.

研究実績

[1] 木村秋斗, 平川翼, 山下隆義, 藤吉弘亘, “Replution Loss を用いた Single Stage Headless 構造による遠方歩行者検出”, 画像センシング技術研究会, 2020.

[2] 木村秋斗, 長内淳樹, 平川翼, 山下隆義, 藤吉弘亘, “Pixel-wise-Attention による Point 型歩行者検出の判断根拠の可視化”, 自動車技術会学術大会, 2021.