Grid-wise-attentionによる物体検出の視覚的説明

木村秋斗† 早川和希† 森巧磨†† 長内淳樹†† 平川翼† 山下隆義† 藤吉弘亘†
†中部大学 ††(株) 本田技術研究所
E-mail: kazu@mprg.cs.chubu.ac.jp

1 はじめに

歩行者検出は，画像上に存在する歩行者の位置を検出する技術である．そのため，自動運転では前方にいる歩行者を検出し，車を停止させるなど，自動運転の安全性を確保するために必要な技術であると言える．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．

従来の物体検出として，Convolutional Newral Network(CNN)[1] による物体検出を用いて行う手法が多数提案されている [2][3]．中でも，Faster R-CNN[4] が，事前に定義した矩形のテンプレートのセットを用いることで物体の位置推定とクラス分類を 1 つのネットワークで行うことを実現したことで，高速化かつ高精度な物体検出手法が提案された [5][6]．しかし，従来の物体検出手法は，ネットワークが物体を検出すると判断した根拠が不明であるという問題がある．

本研究は，物体検出の判断根拠を示すことが可能な手法の実現を目的とし，画像をグリッド状に分割した各地点に対して，アテンションマップを獲得する手法を提案する．提案手法は，画像上に存在する複数の物体に対し，それぞれの検出する地点に対するアテンションマップを取得できるため，各対象の検出の判断根拠を示すことができる. また，提案手法のクラス推定と位置推定を段階的に行い，クラス推定に用いる特徴を基にアテンションマップを求めることで，クラスらしさを考慮した物体検出の判断根拠となる．そして，アテンションマップによる重み付けを行い位置の推定をすることで物体の注視領域を考慮した検出を行う．本稿では，CityPersons データセットを用いた歩行者検出の精度比較から本手法の有効性を調査し，取得した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調査する．

2 関連研究

2.1 物体検出

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．two-stage 型の物体検出手法は物体領域の推定を行った後，クラス分類を行う．中でも，Fast R-CNN[7] や Faster R-CNN などの歩行者検出手法は高速で高精度な歩行者検出を実現している．

Faster R-CNN は Region Proposal Network(RPN) により物体領域の推定を行う．RPN は，CNN が取得した特徴マップに対して，各地点で事前に定義した数種類の異なるサイズやアスペクト比の矩形のテンプレートのセット (アンカー) を使い， 特徴を基にアンカー毎の物体らしさを示す確率とアンカーからの物体位置の座標を求める．RPN により，物体領域の推定とクラス分類を単一のネットワークで行うことができるため，より高速な検出ができる．

one-stage 型の歩行者検出手法は歩行者領域の推定と歩行者かどうかの分類を単一のネットワークで行う．また，処理も一度に行うことができるため，one-stage 型の物体検出手法は two-stage 型と比較して，高い精度を達成しつつも高速化されており，物体検出の代表的なアプローチとなっている．one-stage 型の手法には You Only Look Once(YOLO)[8] や Single shot multibox detector(SSD)[9] がある．

YOLO は入力画像を一定間隔のグリッド状に分割し，グリッド内の領域ごとに，クラス分類と物体領域の推定を行うことで高速な物体検出を行うことができる．しかし，YOLO は分割するグリッドの間隔は事前に定義したもので一定となるため，広範囲のスケールの検出に対応することができないという問題がある．

YOLO を改良した You Only Look Once v3(YOLOv3)[10] は，分類と領域の推定を 3 つの異なる間隔で分割されたグリッド内の領域でそれぞれ行うことで，広範囲のスケールの検出に対応することができる．結果，YOLO よりも高精度に検出ができる．

Center and Scale Prediction(CSP)[11] 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．CSP は，Faster R-CNN や YOLOv3 と異なり，アンカーを使用せずに物体を検出できる．これにより，アンカーのハイパーパラメータによる初期定義に依存しない小さい物体やオクルージョンに頑健な検出ができる．

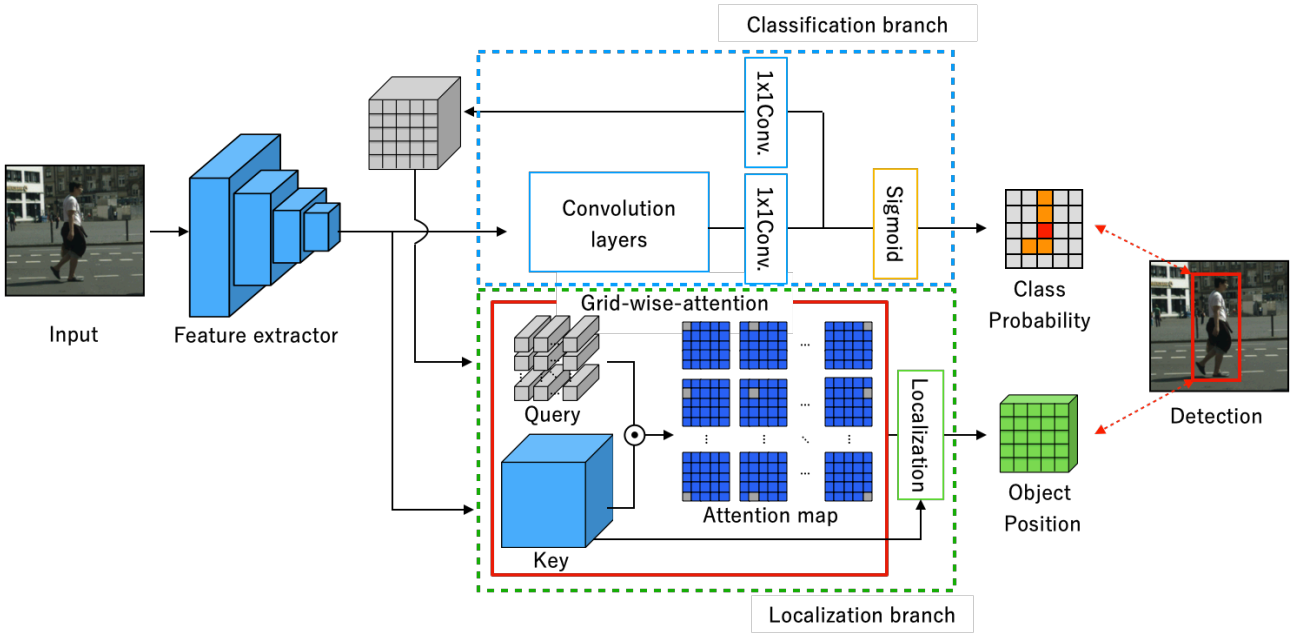

図 1　提案手法のネットワーク構造

2.2　判断根拠の可視化

深層学習モデルの判断根拠を示すための注視領域の視覚的説明は，ネットワークが取得した特徴から生成されるアテンションマップを可視化することで求めることができる．Attention Branch Network[12] は，推論時にネットワークが注視する領域を Attention 機構により認識処理に活用することで注視領域の可視化と高精度な画像認識を実現している．しかし，物体検出は対象が 1 つの画像上に対し複数存在することもあり，ネットワークから取得した特徴から生成される 1 枚のアテンションマップでは表現が難しい．

DEtection TRansformer(DETR)[13] は，畳み込み処理によって獲得した特徴マップから取得した Self-Attention を可視化することで，特徴マップのそれぞれの位置の特徴に対し，他の特徴がどれだけ関係しているかを表す領域が出せる．Self-Attention は，複数の入力にそれぞれ与えられた，Query，Key，Value の 3 つの変数を用いて式 (1) に示すように求める．

$$\mathrm{Attention}(Q,K,V) = \mathrm{softmax}(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}})V \qquad (1)$$

ここで，Q，K，V はそれぞれ Query，Key，Value，d_k は Query の次元数である．DETR は，取り出した地点にいる対象の物体に対し強く注視するアテンションマップを取得することができる．しかし，取り出した地点が対象の物体の検出に最も寄与した地点であるかが不明であり，物体検出の判断根拠として示すには不十分である．

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．

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，図 2 に示す Classification branch に入力する．Classification branch では，1×1 の畳み込み層，Batch Normalization 処理，ReLU 関数で構成される畳み込み層を用いてより詳細な特徴を取得し，1×1 の畳み込み層，シグモイド関数によって各グリッドに対する物体のクラス確率を出力する．classification branch の 1×1 の畳み込み層によって取得したクラス分類に対し有益な情報を持った特徴マップを Localization branch へと入力する．Localization branch では，図 3 に示すように Grid-wise-attention 機構と Localization 処理によって各グリッドに対する物体の位置を出力する．Grid-wise-attention 機構は，$W \times H$ 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

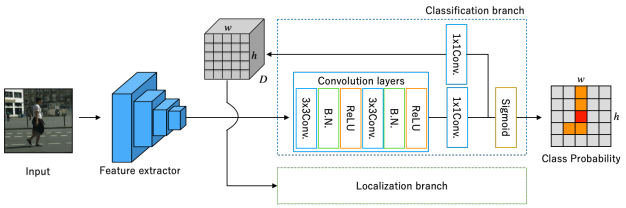

図 2　Classification branch

するアテンションマップ $\mathrm{Attention}_{ij}$ を式 (2) に示す．

$$\mathrm{Attention}_{ij} = \mathrm{softmax}(Q_{ij}K^{\mathrm{T}}) \tag{2}$$

取得される $\mathrm{Attention}_{ij}$ は，Transformer[14] 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．Localization 処理では，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．重み付けされた特徴マップを 1×1 の畳み込み層を用いて特徴を集約し，重みづけしたアテンションマップに対応するグリッドに対しての物体位置を求める．この処理をグリッド毎に行い，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て最も確信度の高いグリッドに対する結果のみ出力する．

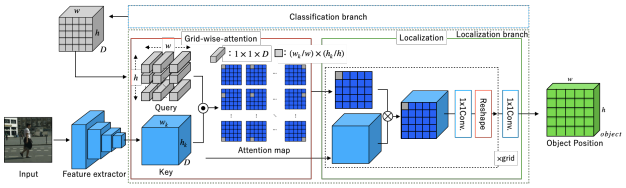

図 3　Localization branch

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[15] 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 (3)，式 (4) に示す．

$$L_{cls} = \frac{1}{N^c}\sum_i -(1-\hat{p}_i)^{\gamma}\log(\hat{p}_i) \tag{3}$$

$$L_{box} = \frac{1}{N^r}\sum_i I(g_i = 1) l_r(b_i, t_i) \tag{4}$$

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．

$$\hat{p}_i = \begin{cases} p_i & \text{if } g_i = 1 \\ 1 - p_i & \text{otherwise} \end{cases} \tag{5}$$

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，図 4 に示すように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．

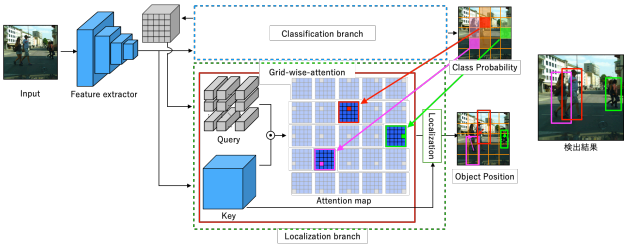

図 4　物体毎のアテンションマップの取得方法

3.4　マルチスケール構造の導入

本手法は，図 5 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．

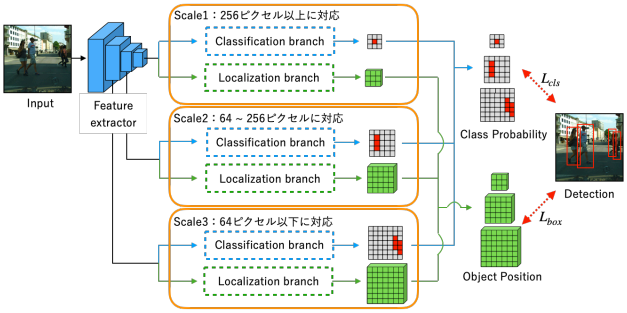

図 5　マルチスケール構造

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．データセットの画像サイズは $2,048 \times 1,024$ ピクセルだが，本実験でネットワークに入力する画像サイズは，メモリサイズの関係上，学習，評価ともに元画像の半分である $1,024 \times 512$ ピクセルとする．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[16] 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[11]	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 6 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．

検出した歩行者のアテンションマップを図 7 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 7(a) から

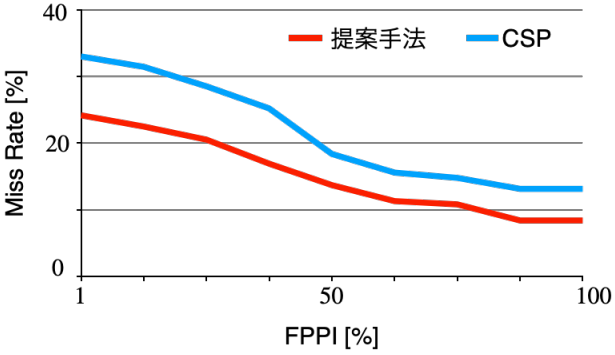

図 3　DET カーブ

(c) のアテンションマップから，歩行者と推測される対象の上半身付近を強く注視していることがわかる．また，図 7(a) 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 7(b)，(c) 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，歩行者の上半身に注目して検出していることがわかる．一方で，図 7(c) 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 7(d) 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ことが影響していることが考えられる．

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上半身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．

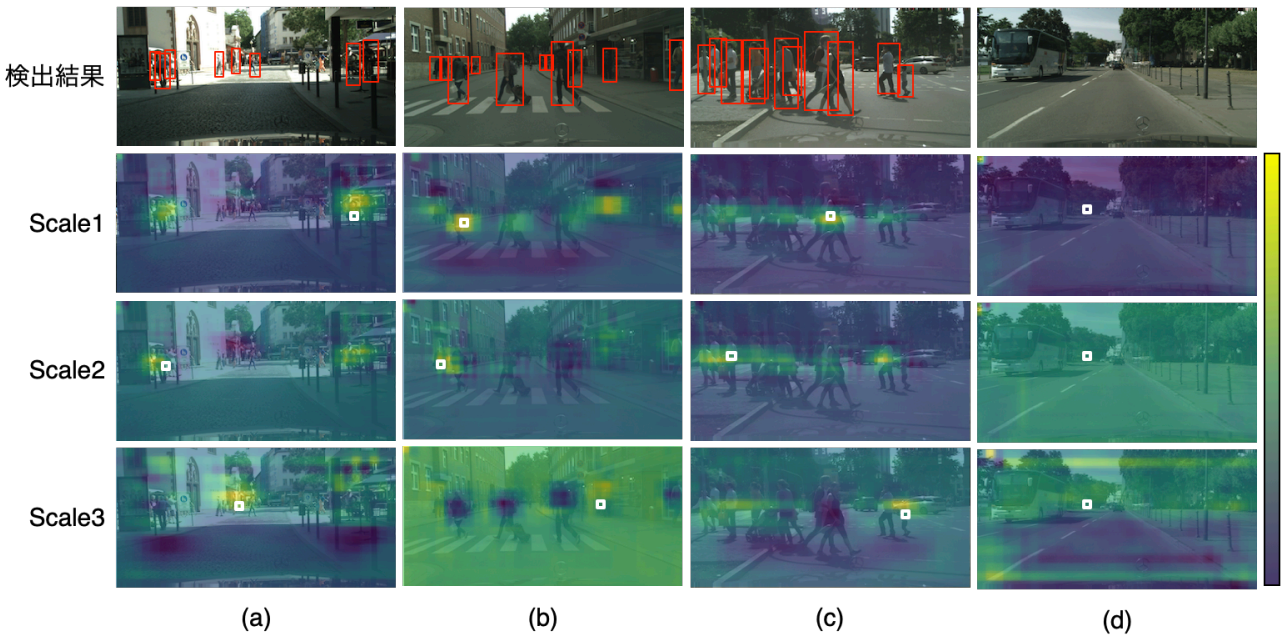

図 4　アテンションマップ可視化結果

参考文献

[1] Bengio Yoshua LeCun Yann, Bottou Léon and Haffner Patrick. Gradient-based learning applied to document recognition. *the IEEE*, 86(11):2278–2324, 1998.

[2] Shaoqing Ren Kaiming He, Xiangyu Zhang and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *European Conference on Computer Vision*, pages 346–361, 2014.

[3] Zhaowei and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *he IEEE Conference on Computer Vision and Pattern Recognition*, pages 6154–6162, 2018.

[4] Girshick Ross Ren Shaoqing, He Kaiming and Sun Jian. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[5] Yongtao Wang Zhi Tang Ying Chen Ling Cai Haibing Ling Qijie Zhao, Tao Sheng. M2det: A single-shot object detector based on multi-level feature pyramid network. In *The Thirty-Third AAAI Conference on Artificial Intelligence*, pages 9259–9266, 2019.

[6] Ruoming Pang Mingxing Tan and Quoc V. Le. Efficientdet: Scalable and efficient object detection. In *The Thirty-Third AAAI Conference on Artificial Intelligence*, pages 10781–10790, 2020.

[7] Ross Girshick. Fast r-cnn. In *the IEEE International Conference on Computer Vision*, pages 1440–1448, 2015.

[8] Girshick Ross Redmon Joseph, Divvala Santosh and Farhadi Ali. You only look once: Unified, real-time object detection. In *the IEEE Conference on Computer Vision and Pattern Recognition*, pages 779–788, 2016.

[9] Erhan Dumitru Szegedy Christian Reed Scott Fu Cheng-Yang Liu Wei, Anguelov Dragomir and Berg Alexander C. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*, pages 21–37, 2016.

[10] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. volume abs/1804.02767, 2018.

[11] Wiqiang Ren Weidong Hu Wei Liu, Shengcai Liao and Yian Yu. High-level semantic feature detection: A new perspective for pedestrian detection. In *the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5187–5196, 2019.

[12] Takayoshi Yamashita Hiroshi Fukui, Tsubasa Hirakawa and Hironobu Fujiyoshi. Attention branch network: Learning of attention mechanism for visual explanation. In *the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10705–10714, 2019.

[13] Gabriel Synnaeve Nicolas Usunier Alexander Kirillov Nicolas Carion, Francisco Massa and Sergey

Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229, 2020.

[14] Niki Parmar Jakob Uszkoreit Llion Jones Aidan N Gomez LukaszKaiser Ashish Vaswani, Noam Shazeer and IlliaPolosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[15] Ross Girshick Kaiming He Tsung-Yi Lin, Priya Goyal and Piotr Dollar. Focal loss for dense object detection. In *the IEEE International Conference on Computer Vision*, pages 2980–2988, 2017.

[16] Bernt Schiele Piotr Dollar, Christian Wojek and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. In *IEEE transactions on pattern analysis and machine intelligence 34(4)*, pages 743–761, 2012.