情報処理学会論文誌 Vol.62 No.12 2069–2078 (Dec. 2021)

広域特徴と狭域特徴を用いた画像情報の部分欠損に頑健な歩行者検出

王 彧[1,a)] 小平 美沙季[2,b)] 加藤 ジェーン[1,c)]

受付日 2021年2月18日, 採録日 2021年9月9日

概要：本論文では，車載カメラ画像に対する歩行者検出の精度改善を目的として，広域特徴と狭域特徴を用いた検出器の構造を提案する．歩行者検出タスクの課題として，(1) オクルージョン発生時，(2) 夜間における精度低下があげられる．(1)，(2) の状況は，いずれも歩行者の見かけに関する情報の欠損を招く点で共通している．我々は，既存手法が単一かつ大きいサイズの受容野（特徴量を計算する過程で，入力画像に対して参照される範囲）のみを持つ CNN に依存することに問題があると考える．したがって，本研究では，受容野に着目し，異なるサイズの様々な受容野を兼ね備える CNN を構築することにより，問題の改善を目指す．具体的には，画像上の広域的な特徴を表現可能な CNN である VggNet（受容野が大きい）と，狭域的な特徴を表現可能な CNN である BagNet（受容野が小さい）の双方を R-CNN の構造内に取り入れ，それぞれの特徴量を組み合わせて活用する．複数のデータセットにおいて評価実験を行った結果，提案手法により，見かけの情報が部分的に欠損した歩行者に対する特徴表現能力の強化および歩行者検出精度の改善が確認された．

キーワード：歩行者検出，深層学習，畳み込みニューラルネットワーク，受容野，広域・狭域特徴

Occlusion Robust Pedestrian Detection with Global and Local Features

Yu Wang[1,a)] Misaki Kodaira[2,b)] Jien Kato[1,c)]

Received: February 18, 2021, Accepted: September 9, 2021

Abstract: We propose a novel pedestrian detector which utilizes both global features and local features. Traditional CNN based pedestrian detectors have difficulties in dealing with scenes of severe occlusions or in the night time, mainly due to the information loss that caused by using a single CNN with a relatively large receptive field. In the proposed approach, we utilize both the VggNet and the BagNet, which are designed to have large and small receptive fields respectively, to capture information from the input images in different scales. We evaluated the proposed approach on multiple pedestrian detection datasets and confirmed its effectiveness.

Keywords: pedestrian detection, deep learning, CNN, receptive field, BagNet, local/global features

1. はじめに

歩行者の交通事故は，日本において深刻な社会問題となっている．平成 30 年度の警視庁の統計 [1] によると，状況別[*1]での死者数の構成率は「歩行中」が約 36%と最も高い．このため，近年では事故を未然に防ぐことを目的とした予防安全技術が注目を集めている．歩行者検出技術はその重要な基盤要素の 1 つである．

歩行者検出技術はすでに段階的に実用化が進んでいる

[1] 立命館大学情報理工学部
Ritsumeikan University, Kusatsu, Shiga, 525–8577 Japan
[2] 名古屋大学大学院情報学研究科
Nagoya University, Nagoya, Aichi, 464–8601 Japan
a) ywang@nagoya-u.jp
b) misaki@nagoya-u.jp
c) jien@nagoya-u.jp

[*1] 「自動車乗車中」,「自動二輪車乗車中」,「原付乗車中」,「自転車乗車中」,「歩行中」の 5 通りの状況が考慮されている．

が，依然として検出エラーの問題が残っている．特に検出エラーが起こりやすい例としては，(1) 他物体によるオクルージョン，(2) 夜間の白飛び・黒潰れなどにより，情報の一部が欠損している場合があげられる．検出性能を向上させるためには，このような画像情報の部分的な欠損に対して頑健な検出アルゴリズムを構築することが要求される．歩行者検出に関する近年の研究では，畳み込みニューラルネットワーク（CNN）[2], [3], [4] を用いる手法が主流となっており，認識性能の最高記録を次々と塗り替えている．一方，CNN により計算される特徴量は，上述の (1), (2) のような状況において検出精度の低下を招きやすいという問題点が存在する．一般的に CNN では，入力画像に対して多層にわたるフィルタを適用し特徴量を計算する過程で，参照される範囲（受容野）が広域的に広がる性質がある．ここで，画像中の広域的な範囲を表現する特徴量を「広域特徴」，狭域的な範囲を表現する特徴量を「狭域特徴」と定義するとき，これら 2 種類の特徴量にはそれぞれの利点・欠点が存在する．広域特徴は，歩行者の全体的な形状をとらえやすいが，領域内に映り込んだ他物体の影響を受けやすい．一方，狭域特徴は，変形やオクルージョンに強く，小さな歩行者を表現しやすいが，参照範囲の位置・大きさによっては歩行者の識別が困難である．CNN を用いた従来の歩行者検出手法 [5], [6], [7] では，広域特徴に依存するものがほとんどである．これらの手法は，特徴量の参照範囲に対して歩行者の占める範囲が限定的になり，歩行者らしさを表現することが困難となる．

このような問題意識を用いて本研究では，受容野に着目し，異なるサイズの様々な受容野を兼ね備える CNN を構築することにより，上述 (1), (2) の問題点の改善を目指す．具体的には，広域特徴と狭域特徴を組み合わせ，双方の強みを活かすことのできる歩行者検出器のモデルを提案する．提案モデルは，前段の歩行者の候補領域を抽出する検出器および後段の候補領域を歩行者/背景に分類することで誤検出を排除する分類器から構成される．従来の 2 段式歩行者検出手法と異なり，前段の検出器は，大きい受容野と小さい受容野を持つことで，広域特徴と狭域特徴を同時に抽出する，後段の分類器は，複数の画像サイズおよび畳み込み層から得られた特徴を融和することで，実質的に広域特徴と狭域特徴を同時に抽出する，広域・狭域特徴の双方の強みを活かすことのできるモデル設計を実現する．提案手法を代表的な歩行者データセットである Caltech Pedestrian Dataset [8], [9]，CityPersons Dataset [10] および NightOwls Dataset [11] を用いて評価した結果，上述の (1)，(2) の状況において提案手法の有効性を確認した．

2. 関連研究

本章では，CNN を用いた歩行者検出および CNN のアーキテクチャに関する関連研究について述べ，本研究の位置づけを明らかにする．

CNN を用いた歩行者検出． 画像からの特徴抽出手法として，Haar-like [12] 特徴や HOG [8] 特徴をはじめとする手動で設計される特徴量を活用した手法が一般的であった．一方，深層学習の普及にともない，CNN を用いて画像特徴を抽出する手法が主流となった．CNN では，多層にわたるフィルタを目的のタスクに合わせて自動的に最適化することで，ハイレベルな特徴の抽出が可能である．これにより，手動で特徴量を設計する従来手法と比較して特徴表現の性能が飛躍的に向上した．CNN を用いた歩行者検出手法としては，個別の検出器・分類器を用いた手法 [13], [14], R-CNN ベースの手法 [5], [7], [10], [15], [16], [17], [18], [19], [20], Single Shot ベースの手法 [6], [21], [22], [23] などが存在する．

個別の検出器・分類器を用いた手法では，前段で LDCF [24] などの検出器を用いて歩行者候補領域を抽出し，後段で CNN の分類器を用いて候補領域を歩行者/背景に分類することで誤検出を排除する．具体例として，TA-CNN [13] では，ACF [25] を用いて歩行者候補領域を抽出した後，CNN を用いて歩行者/背景の分類および属性認識を行っている．また，DeepParts [14] では，LDCF を用いて歩行者候補領域を抽出した後，歩行者領域の部位ごとに学習された複数の CNN を用いてそれぞれのスコアを出力し，SVM [26] で最終的な歩行者スコアを算出する．これらの手法では，後段の CNN に工夫を施すことで誤検出を削減したが，前段部分での未検出が性能改善に対するボトルネックとなった．

R-CNN ベースの手法では，Selective Search [27] や Region Proposal Network（RPN）[28] を用いて抽出される歩行者候補領域を，Boosted Tree や全結合層などを通じて歩行者/背景に分類する．このなかでも特に RPN を用いた手法では，候補領域の抽出時に CNN を使用することで検出精度を向上させると同時に，畳み込み処理によって画像のスケールを圧縮したうえで候補領域を走査することで，処理速度の向上を実現している．具体例として，SA-FastRCNN [20] や MS-CNN [18] では，歩行者ごとのスケールの違いを考慮したモデルを R-CNN のフレームワーク中に統合している．また，AdaptedFasterRCNN [10] では，RPN のアンカーのスケールやストライドなどのハイパーパラメータを適切に設定することにより，単純な Faster R-CNN [28] でも比較的高い性能が得られることを示している．その他にも，FasterRCNN+ATT [17] では，CNN の異なるチャンネル特徴量がそれぞれ異なる人体パーツに反応して活性化することを利用して，オクルージョンに対して頑強なモデルを提案している．

一方で，RPN のみを利用した手法も存在する．RPN+BF [5] では，歩行者検出においては Faster R-CNN の前半の RPN を単独で使用することで高い性能が得られることが示されている．また，後半部分のクラス分類器

を Boosted Forest [29] で置き換えることによって，さらに精度を向上させている．同様に，SDS-RCNN [7] においても RPN を単独で使用し，後半部分のクラス分類器として VggNet を使用している．この手法では，前半・後半双方の VggNet にセグメンテーション層を加えることによって，歩行者の検出を補助している．その他，AR-Ped [16] では，RPN の内部で IoU の閾値を徐々に厳しくしながら歩行者候補領域に対するスコアの推定を行う自己回帰モデルを構築することによって，高い検出性能を実現できている．

Single Shot ベースの手法では，グリッドに沿って検出を行うことにより，1 つの CNN 内の 1 回の処理で歩行者検出を可能としている．これらの手法では，最終的な検出結果を出力するまでのステップ数が少なく，高速な検出が可能である一方，検出精度の面では R-CNN ベースの手法よりも性能が低くなる傾向がある．具体例として，Noh ら [30] は，YOLO [31] や SSD [32] をはじめとする 1 段階の物体検出器に対し，オクルージョンや Hard Negative を考慮して出力テンソルを更新する End-to-end で学習可能なシステムを導入することによって，歩行者検出での精度改善を図っている．また，TLL-TFA [23] では，ResNet を用いて入力画像から特徴マップを計算した後，それらに対して Deconvolution を適用し，出力される最終的な特徴マップをもとに歩行者の頭上から足先までを結ぶ線分を推定する．この手法では，歩行者を囲む矩形領域ではなく，歩行者に対応する線分を学習・推定することによって，アノテーションの曖昧さを軽減し，検出精度の改善を図っている．

本論文では比較的に高い検出性能を実現できる R-CNN ベースの検出構造を利用する．R-CNN ベースのほとんどの既存手法 [5], [7] では，前段も後段も「広域特徴」のみが用いられるが，本研究では，前段について，特徴抽出部は，受容野の広い VggNet だけでなく，受容野の狭い BagNet も取り入れ，前者と後者によりそれぞれ抽出された広域特徴と狭域特徴を併用している．また，後段について，入力の歩行者候補領域を複数段階にリサイズし，同一受容野の VggNet に入力することで，疑似的に特徴マップの広域/狭域性を持たせている．これらの工夫により，「広域特徴」だけではなく，「狭域特徴」も利用することになり，検出精度の向上をもたらした．

CNN のアーキテクチャ．物体検出を行う際に用いられる CNN としては，大規模画像認識データセットである ImageNet [33] 上で高い性能を発揮させる目的で提案された画像認識用 CNN のアーキテクチャがしばしば流用されている．主な CNN のアーキテクチャとしては，AlexNet [2], GoogLeNet [34], VggNet [3], ResNet [4], ResNeXt [35], Xception [36], DenseNet [37], DRN [38], DetNet [39], BagNet [40] など様々な種類が存在する．

AlexNet は，少ない層数で構成され，各層が 1 列に連なった単純な構造となっている．ゆえに高速な計算が可能である一方，特徴抽出の性能は他の CNN と比較して劣る傾向にある．GoogLeNet では，Inception モジュールを定義し，これらを組み合わせて 1 つの CNN を構成することにより，多様な特徴の表現を実現している．VggNet では，AlexNet よりもスケールの小さなフィルタがより多く連なった構造をとっており，特徴抽出の性能が高められている．ResNet では，Shortcut Connection の導入によって勾配消失の問題を解消し，さらに層の深い構造を実現している．ResNeXt や Xception では，従来の畳み込みを Grouped Convolution に置き換えることにより，パラメータ数の削減と特徴表現能力の向上を同時に実現している．DenseNet では，ResNet の Shortcut Connection を発展させ，特徴マップのスケールが同じであるすべての層を結合させた構造をとっている．これにより，層間の情報伝達を最大化し，高い特徴表現能力を実現している．DRN や DetNet では，Dilation を適用し，特徴マップの解像度の低下を防ぐことによって，画像認識やセグメンテーション，物体検出の精度向上を図っている．以上のような CNN で計算される特徴マップは，受容野が非常に大きい点で共通している．

一方，BagNet では，ResNet と同様の Shortcut Connection を活用した構造を採用しつつ，畳み込みフィルタのサイズ，ストライドを小さくすることにより，特徴マップの受容野を非常に小さく抑えている．このアーキテクチャでは，画像中の限定的な範囲に対して繰り返し畳み込み処理が行われるため，高レベルの狭域特徴が計算可能である．

本論文では受容野が大きい CNN である VggNet と，受容野が小さい CNN である BagNet を組み合わせ，歩行者検出に適した新しいアーキテクチャを提案する．

3. 問題設定

受容野の大きさは，CNN 各層のフィルタサイズおよびストライドによって決定される．ある CNN において，フィルタサイズおよびストライドの縦，横がつねに等しいという前提を置き，l 番目の層のフィルタサイズを k_l，ストライドを s_l，出力される特徴マップの受容野の大きさを r_l とする．このとき r_{l-1} は式 (1) にしたがって計算される [41].

$$r_{l-1} = s_l \cdot r_l + (k_l - s_l) \tag{1}$$

また，式 (1) から，入力画像上での受容野の大きさ r_0 は式 (2) のように導出される．

$$r_0 = \sum_{l=1}^{L} \left((k_l - 1) \prod_{i=1}^{l-1} s_i \right) + 1 \tag{2}$$

以下，本論文では，r_0 が大きい特徴量を広域特徴，小さい特徴量を狭域特徴と呼び，区別する．

歩行者検出タスクにおける広域特徴の利点は，歩行者の全身の形状をとらえることができ，全体的な特徴をふまえて歩行者を識別しやすい点である．このため**図 1**（左）の

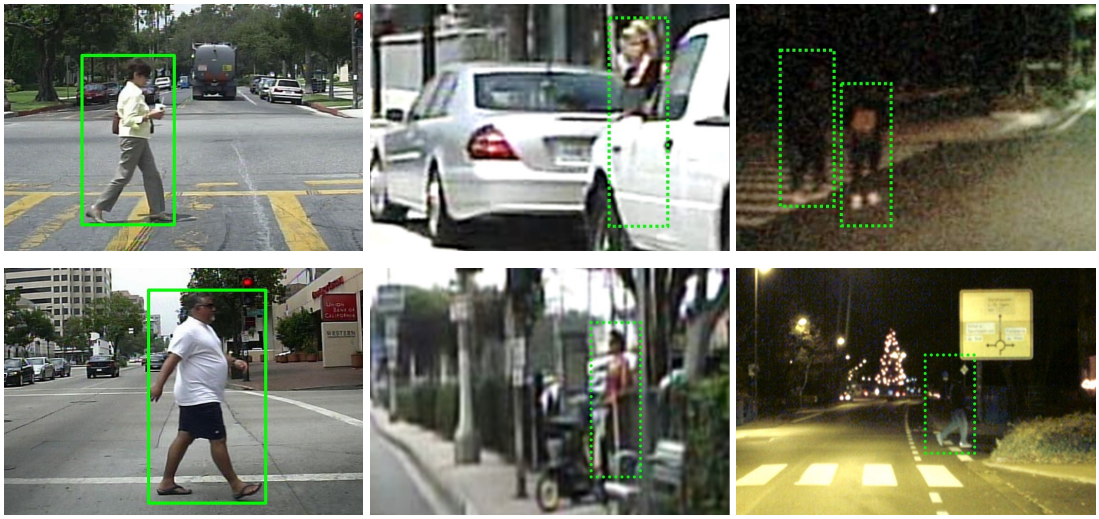

図 1 （左）：広域特徴で表現しやすい歩行者の例．自車との距離が近く，全身が鮮明に写っている．（中）：広域特徴で表現しにくい歩行者の例．歩行者が他物体の背後に隠れており，確認できる範囲が限定的である．（右）：広域特徴で表現しにくい歩行者の例．夜間で周囲が暗く，歩行者が部分的に背景に同化している

Fig. 1 (Left): Examples of pedestrians that can be well represented using global features. They are usually close to the onboard camera with the whole body been well captured. (Middle): Examples of pedestrians that can not be well represented using global features. They are partially occluded. (Right): Examples of pedestrians that can not be well represented using global features. Because the surroundings are dark at nights, the pedestrians are partially assimilated into the background.

ように，昼間で歩行者が自車の近くに存在し，おおむね全身が見えている場合であれば効果的に機能する．一方で，広域特徴の欠点は，参照領域内に写りこんだ他物体の影響を受けやすい点である．このため図 1（中）のように，身体の一部が他物体の背後に隠れている場合には，参照領域内で歩行者が写っている面積が相対的に狭くなり，歩行者の特徴を表現しにくくなる．同様に，夜間は周囲が暗いため，図 1（右）のように車のヘッドライトなどで照らされている身体の一部のみを見ることができ，他の部分は背景に同化してしまうケースが多いことから，広域特徴では特徴を表現しにくく，検出性能の低下を招きやすい．

以上の課題をふまえ，本論文では，広域特徴，狭域特徴の双方の強みを活かすことによって，検出性能を向上させる手法を提案する．

4. 提案手法

提案手法では 2 段式の検出構造を採用し，前段の Multi-scale RPN で歩行者候補領域の抽出，後段の Multi-scale BCN で歩行者/背景の分類を行う（**図 2**）．

4.1 Multi-scale RPN による歩行者候補領域の抽出

前段の Multi-scale RPN では，既存手法 SDS-RCNN [7] の RPN（Region Proposal Network）の構造をベースとする．この RPN は通常の RPN と同様に歩行者候補領域のバウンディングボックスおよび信頼度スコアを出力する候補領域抽出層（proposal）のほかに，各ピクセルごとの歩行者・背景の確率を表現するセグメンテーションマスクを出力するセグメンテーション層（seg）を加えることで，歩行者領域の検出をより高精度で実現している．また，この RPN では特徴抽出をする際に VggNet のみを使用するが，狭域特徴を導入するため，提案手法の Multi-scale RPN では，VggNet および BagNet の 2 種類の CNN から特徴抽出を行い，それぞれの特徴マップをチャンネル方向に連結して使用する．これら 2 種類の CNN は，受容野の観点から大きく性質が異なる．

VggNet は 5 つの畳み込みブロックおよび 2 つの全結合層から構成されており，各畳み込みブロックの内部では，フィルタサイズ 3×3，ストライド 1 の畳み込み層 2〜4 層を適用した後，ウィンドウサイズ 2×2，ストライド 2 のマックスプーリング層によりダウンサンプリングを行う．ネットワークの枝分かれはなく，すべての層が直鎖状に連なっている．本論文の提案手法では VggNet16 の全結合層を取り除き，特徴抽出器として使用する．また，特徴マップの縦横を BagNet に合わせるため，pool4，pool5 を排除する．この特徴抽出器から出力される特徴マップの受容野 r_0 は 140×140 であり，車載カメラ画像に適応すると，多くの場合歩行者の全身がカバーされる．

一方，BagNet は ResNet50 をベースに開発されたモデ

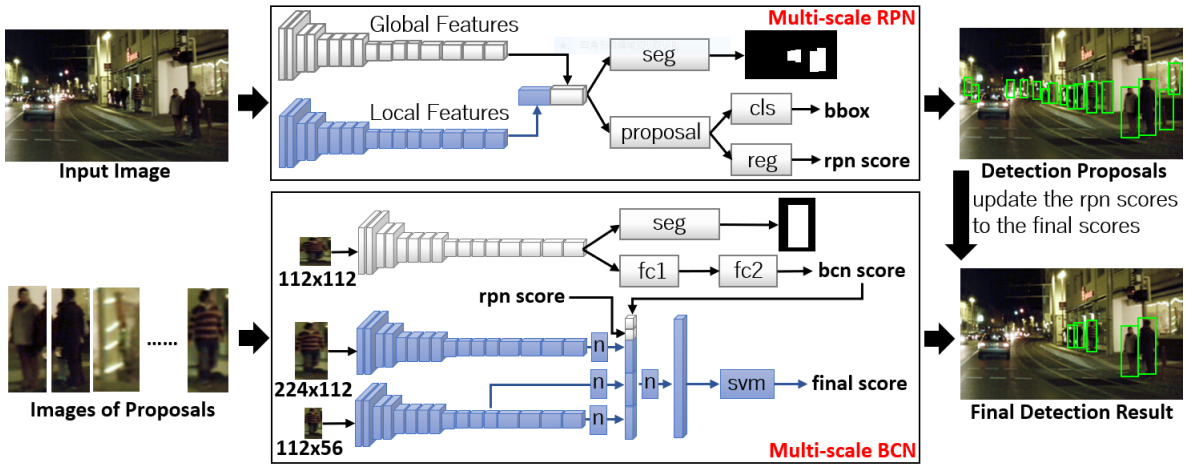

図 2 提案手法の全体像．図中の seg はセグメンテーション層，cls はクラス分類層，reg はバウンディングボックスの回帰層，fc1，fc2 は全結合層，n は正規化処理を表す．また，水色の処理ブロックは SDS-RCNN と提案手法の差分を示す

Fig. 2 Overview of the proposed method. In the figure, "seg" means segmentation layer, "cls" means classification layer, "reg" means bounding box regression layer, "fc1" and "fc2" are fully connected layers, and "n" means normalization layer. Processing blocks in light blue show the difference comparing to the SDS-RCNN.

ルであり，BagNet9，BagNet17，BagNet33 の 3 種類が存在する．これらのモデルは ResNet50 とほとんど同様の構造となっているが，パディング操作をなくしたうえ，ほとんどの畳み込み層のフィルタサイズを 3×3 から 1×1 に変更することにより，受容野が小さく抑えられている．本論文の提案手法では BagNet33 の global average pooling 層および全結合層を取り除き，特徴抽出器として使用する．この特徴抽出器から出力される特徴マップの受容野 r_0 は 33×33 であり，車載カメラ画像に適応すると，多くの場合歩行者の身体のパーツがカバーされる．

4.2 Multi-scale BCN による歩行者背景の分類

後段の Multi-scale BCN では，既存手法 SDS-RCNN の BCN（Binary Classification Network）の構造をベースとする．既存手法では，入力画像から特徴を抽出する際に単一の VggNet のみを使用するが，提案手法では，先行研究 [42] の評価実験に基づき，VggNet に下記 (1)〜(3) 項の改良を加えることでマルチスケールの特徴抽出を実現する．

(1) 入力画像サイズ

提案手法では 2 つの VggNet を使用し，候補領域画像を一方では 224×112 に，もう一方では 112×56 にリサイズして入力し，特徴抽出を行う．

元画像における参照範囲の大きさ a は，画像のスケーリング率と受容野の一辺の長さ r_0 に依存する．ここで，一辺の長さ h の正方形の入力画像を考える．この画像をそのまま CNN に入力する場合，$a=r_0$ が成り立つ．次に，一辺の長さを h' にリサイズして CNN に入力する場合を考える．このとき，$a'=\frac{h}{h'}r_0$ となる．したがって，参照範囲 a は入力画像を拡大するほど狭域的に，縮小するほど広域的になるといえる．

(2) 特徴抽出層

提案手法では，入力画像サイズ 224×112 の VggNet の pool5，112×56 の VggNet の pool4，pool5 の計 3 カ所から特徴抽出を行う．

CNN では，入力に近いほど受容野が小さく，エッジやテクスチャなどのローレベルな特徴が表現される．また，出力に近い層ほど受容野が大きく，「歩行者らしさ」を表現するハイレベルな特徴が表現される．

(3) 正規化および複数特徴ベクトルの連結

抽出した特徴マップの各ピクセルを 1 列に並べてベクトル化し，式 (3) で示す符号付き平方根および L2 正規化を適用する．ただし x_i は正規化適用前，z_i は正規化適用後の特徴ベクトルの i 番目の要素を表す．正規化によりデータが平滑化され，特定の要素の影響が過度に強まる現象を防ぐことができる．また，複数の特徴ベクトルを連結する際に，各々の影響力を平等化することができる．

$$y_i=\mathrm{sign}(x_i)\sqrt{|x_i|}\,,\qquad z_i=\frac{y_i}{\sqrt{\sum_{i=1}^{d}y_i^2}} \tag{3}$$

提案手法では，抽出・正規化した特徴ベクトルどうしを連結し，SVM により歩行者/背景に分類する．参照範囲が実質的に異なる複数の広域特徴および狭域特徴を組み合わせて使用することにより，特徴の表現力を強化する．

5. 実験設定

5.1 データセット

提案手法のパフォーマンスについて詳細に分析するため，異なる 3 種類のデータセット上で実験を行った．

5.1.1 Caltech Pedestrian Dataset

Caltech Pedestrian Dataset [8] は車載カメラで撮影された走行シーンから構成されており，約 35 万の歩行者アノテーションを含んでいる．画像の解像度は 640 × 480 である．学習用画像の枚数は 42,782 枚，評価用画像の枚数は 4,024 枚である．本研究では，条件の異なるサブセット Reasonable（高さ 50 ピクセル以上で，全身の 35%未満のオクルージョンがあるか，オクルージョンのない歩行者）と Heavy Occlusion（高さ 50 ピクセル以上で，全身の 20%以上 65%以下のオクルージョンがある歩行者）で実験を行い，提案手法の長所・短所を明確化する．

5.1.2 CityPersons Dataset

CityPersons Dataset [10] は車載カメラで撮影された走行シーンから構成されており，約 3.5 万の歩行者アノテーションを含んでいる．Caltech Pedestrian Dataset よりも画像 1 枚あたりの歩行者の人数が多く，歩行者のオクルージョンが多く発生しているという特徴がある．画像の解像度は 2,048 × 1,024 である．学習用画像の枚数は 2,975 枚，評価用画像の枚数は 500 枚である．

5.1.3 NightOwls Dataset

NightOwls Dataset [11] は車載カメラで撮影された走行シーンから構成されており，約 4.2 万枚の歩行者アノテーションを含んでいる．夜間に撮影されたシーンのみで構成されている点が特徴である．画像の解像度は 1,024 × 640 である．学習用画像の枚数は 43,344 枚，評価用画像の枚数は 1,725 枚である．

5.2 評価指標

検出精度の評価には，文献 [9] で提唱されている対数平均ミス率を用いる．対数平均ミス率 MR は，FPPI（False Positives Per Image）が $[10^{-2}, 10^{0}]$ の範囲で等間隔な 9 点におけるミス率 mr の平均をとることにより算出される．式 (4) に計算式を示す．ただし $n = 9$ である．

$$\mathrm{MR} = \exp\left[\frac{1}{n}\sum_{i=1}^{n}\ln mr_i\right] \tag{4}$$

5.3 詳細設定

実験では，(1) 前段の RPN のみを使用した場合，(2) 前段の RPN と後段の BCN の両方を使用した場合の 2 通りについて，比較手法と提案手法の検出精度を検証する．なお，比較手法は，提案手法で一部の構造を利用した既存手法の SDS-RCNN とする．前段のみ使用した実験では，バックボーンの CNN として，a) VggNet を使用した Global RPN（SDS-RCNN の RPN に相当），b) BagNet を使用した Local RPN，および，c) VggNet と BagNet を併用した Multi-scale RPN（提案手法の RPN）の 3 通りを比較する．前段と後段両方を使用した実験では元の SDS-RCNN（Global RPN + BCN），前段のみ改良した SDS-RCNN+（Multi-scale RPN + BCN），前段と後段両方を改良した提案手法（Multi-scale RPN + Multi-scale BCN）の 3 通りを比較する．それによって，広域・狭域のそれぞれの特徴量とそれらを組み合わせた特徴量の有用性を検証する．

RPN 学習時のハイパーパラメータの設定は，SDS-RCNN の設定に従う．ただし，メモリ容量の不足を防ぐため，入力画像のスケールは，Caltech Pedestrian Dataset に関して原画像の 0.8 倍，CityPersons Dataset に関して原画像の 0.3 倍，NightOwls Dataset に関して原画像の 0.5 倍にそれぞれ圧縮して実験に用いる．Multi-scale BCN を構成する各 CNN の学習では，ImageNet で事前学習済み VggNet19 を利用し，入力画像から切り出した ground truth を歩行者，RPN を用いて切り出した歩行者候補領域のうち，ground truth との IoU が 0.5 未満のものを背景として 2 クラス分類の学習を行う．学習時のハイパーパラメータは，ドロップアウト率 0.5，バッチサイズ 128 とする．学習率とエポック数は，はじめに全結合層を学習率 0.001 で 10 エポック学習させた後，全体を 0.0001 で 20 エポック，0.00001 で 20 エポック，0.000001 で 20 エポック学習させる．

6. 実験結果

6.1 定量評価

本節では，対数平均ミス率による定量的な評価を行う．**図 3** の左から右への順に，Caltech Pedestrian Dataset の Reasonable サブセット，Caltech Pedestrian Dataset の Heavy Occlusion サブセット，CityPersons Dataset および NightOwls Dataset 上での評価結果を掲載する．上の行は前段の RPN のみを使用した場合の実験結果，下の行は前段と後段を併用した場合の実験結果となる．

Caltech Pedestrian Dataset の Reasonable サブセットでは，Global RPN が 15.33%，Local RPN が 37.22%，Multi-scale RPN が 14.66%，SDS-RCNN が 10.31%，SDS RCNN+が 9.47%，提案手法が 9.15%となっている．この結果から，前段のみを使用の場合でも，前段と後段を併用した場合でも，広域・狭域特徴の組合せは有効であることが分かる．また，提案手法がオクルージョンの少ない歩行者に対しても有効であることが確認できる．

Caltech Pedestrian Dataset の Heavy Occlusion サブセットでは，Global RPN が 78.58%，Local RPN が 77.66%，Multi-scale RPN が 77.52%，SDS-RCNN が 62.68%，SDS-RCNN+が 62.52%，提案手法が 59.30%となっている．この結果から，オクルージョンの多い歩行者に対しては，前段のみを使用した場合において，狭域特徴は広域特徴よりも有効であることが分かる．また，提案手法はオクルージョンの多い画像に対して特に有効であることも確認できる．さらに，前段と後段を併用した場合，提案手法において，後段の Multi-scale BCN は前段の Multi-scale RPN よ

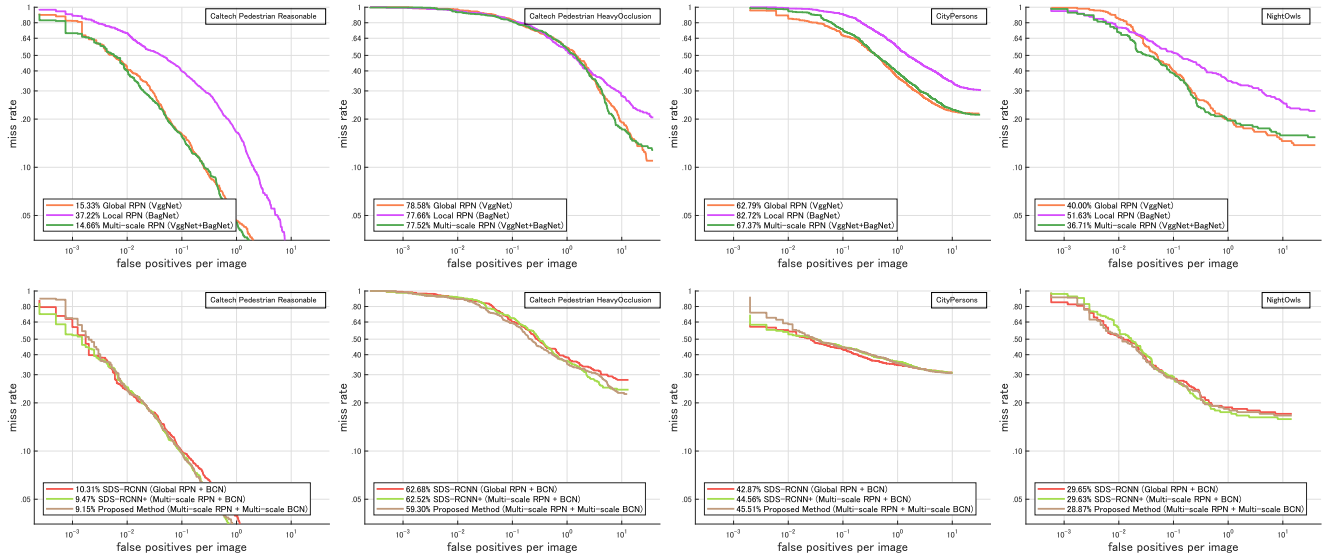

図 3 左から右への順に，Caltech Pedestrian Dataset の Reasonable サブセット，Caltech Pedestrian Dataset の Heavy Occlusion サブセット，CityPersons Dataset および NightOwls Dataset における FPPI-MR グラフを示す．上の行は前段の RPN のみを使用した場合の実験結果，下の行は前段と後段を併用した場合の実験結果である

Fig. 3 From left to right, the subfigures show the FFPI-MR curves of the evaluated approaches on the Caltech Pedestrian Dataset Reasonable Subset, the Caltech Pedestrian Dataset Heavy Occlusion Subset, the CityPersons Dataset and the NightOwls Dataset, respectively. From top to down, the subfigures show results of approaches using RPN only, and results of approaches using both RPN and BCN, respectively.

り効果的であることも明らかである．

一方，CityPersons Dataset では，Global RPN が 62.79%，Local RPN が 82.72%，Multi-scale RPN が 67.37%，SDS-RCNN が 42.87%，SDS-RCNN+が 44.56%，提案手法が 45.51%となっている．いずれの場合においても，提案手法がそれぞれの比較手法より若干精度が低いことが分かる．その原因として，次の 2 点が考えられる．(1) CityPersons は歩行者同士の重なりが頻繁に発生しており，狭域特徴では検出された歩行者のパーツがどの歩行者に属するものかは判断しにくい．(2) 学習画像数は 2,975 枚だけであり，Caltech Pedestrian Dataset の 42,782 枚や NightOwls Dataset の 43,344 枚よりはるかに少ないため，学習データに対して過学習を起こし，評価データに対する汎化能力が低下してしまう．

NightOwls Dataset では，Global RPN が 40.00%，Local RPN が 51.63%，Multi-scale RPN が 36.71%，SDS-RCNN が 29.65%，SDS-RCNN+が 29.63%，提案手法が 28.87%となっている．この結果から，狭域特徴単独での精度は低いものの，広域・狭域特徴の組合せは効果的であることが分かる．また，前段と後段を併用した場合，後段の Multi-scale BCN は前段の Multi-scale RPN よりも大きく検出精度の向上に貢献した．提案手法が夜間の歩行者に対して特に有効であることが確認できる．

6.2 定性評価

本節では，検出結果の可視化による定性的な評価を行う．**図 4**，**図 5** に，各データセットにおける比較手法（SDS-RCNN）と提案手法の前段のみを使用した場合，および前段と後段を併用した場合の検出結果の可視化結果を示す．図中の緑色枠は正しく検出された歩行者，赤色枠は誤検出，青色枠は未検出をそれぞれ表す．CityPersons Dataset においては，提案手法が歩行者の一部あるいは他物体を歩行者と誤検出しやすいことが確認できるが，Caltech Pedestrian Dataset，NightOwls Dataset においては，提案手法は比較手法よりも未検出・誤検出が少ないことが確認できる．

7. おわり

7.1 本論文のまとめ

本論文では，CNN 畳み込み層の重なりによる受容野の拡大という視点から，視覚情報の欠損で生じる歩行者検出エラーの問題を取り上げ，広域・狭域特徴を用いた 2 段式歩行者検出手法を提案した．従来の検出手法と異なり，前段の検出器には，VggNet に加えて受容野の抑制機能を持つ BagNet を導入し，後段の分類器には，異なるスケールの画像特徴を用いることにより，広域特徴，狭域特徴の双方の強みを活かすことのできるモデル設計を実現した．評価実験の結果，提案手法ではオクルージョンのある歩行者や遠くにいる歩行者，夜間の歩行者に対する検出精度が向

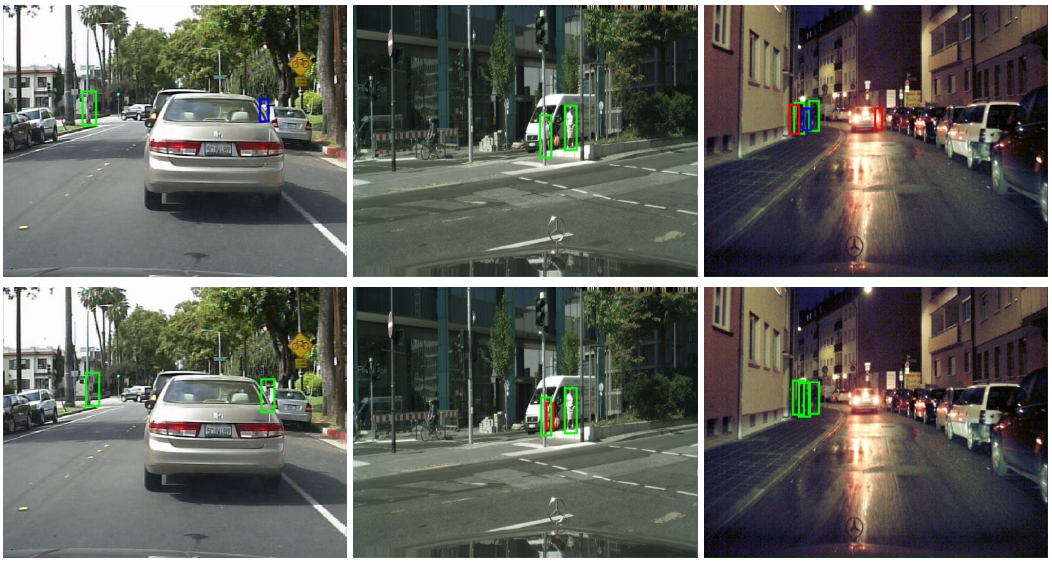

図 4 Caltech Pedestrian Dataset（左），CityPersons Dataset（中）および NightOwls Dataset（右）における SDS-RCNN の Global RPN（上）および提案手法の Multi-scale RPN（下）を用いた検出結果の可視化

Fig. 4 Visualization of detection results produced by the Global RPN (top) and the Multi-scale RPN (down), on the Caltech Pedestrian Dataset (left), the CityPersons Dataset (middle) and the NightOwls Dataset (right).

図 5 Caltech Pedestrian Dataset（左），CityPersons Dataset（中）および NightOwls Dataset（右）における SDS-RCNN（上）および提案手法（Multi-scale RPN + Multi-scale BCN）（下）を用いた検出結果の可視化

Fig. 5 Visualization of detection results produced by the SDS-RCNN (top) and the proposed approach (down), on the Caltech Pedestrian Dataset (left), the CityPersons Dataset (middle) and the NightOwls Dataset (right).

上することが確認された．

したがって，本論文の貢献は下記のとおりである．

- 広域・狭域特徴の双方を活用した独特のネットワーク構造を提案した．
- 提案手法を代表的な歩行者データセット上で性能評価を行い，オクルージョンや夜間撮影により生じる視覚情報の欠損に対して提案手法の有効性を実証した．

7.2 今後の課題

提案手法では，広域・狭域特徴を抽出するため，構造が複雑になり，モデルのパラメータ数も大幅に増えた．ベースラインの SDS-RCNN のパラメータ数（約 0.8 億）と比べて，提案手法はその 1.5 倍以上（約 1.3 億）になり，必要とされるメモリ容量や計算コストも増大してしまう．このような課題を解決するためのアプローチとして，異なる

CNN どうしでパラメータを共有し，必要なメモリ容量や計算コストを削減する工夫が求められる．また，CityPersons Dataset において，学習画像は 2,795 枚しかなく，他のデータセットよりかなり少ないという問題点もある．この問題点に関して，学習画像を増やすか他のデータセットを用いた検証をさらに行う必要があると思われる．

謝辞 本研究の一部は，科研費挑戦的研究（No.18H05394），若手研究（No.20K19831），国立研究開発法人科学技術振興機構の研究成果展開事業「センター・オブ・イノベーションプログラム」の支援によるものである．

参考文献

[1] 交通企画課：平成 30 年における交通死亡事故の特徴等について，入手先 ⟨https://www.npa.go.jp/publications/statistics/koutsuu/jiko/H30sibou_3set.pdf⟩（参照 2019-12-19）.

[2] Krizhevsky, A., Sutskever, I. and Hinton, G.E.: ImageNet classification with deep convolutional neural networks, *Advances in Neural Information Processing Systems*, pp.1097–1105 (2012).

[3] Simonyan, K. and Zisserman, A.: Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556, pp.1–14 (2014).

[4] He, K., Zhang, X., Ren, S. and Sun, J.: Deep residual learning for image recognition, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.770–778 (2016).

[5] Zhang, L., Lin, L., Liang, X. and He, K.: Is faster R-CNN doing well for pedestrian detection?, *European Conference on Computer Vision*, pp.443–457, Springer (2016).

[6] Du, X., El-Khamy, M., Lee, J. and Davis, L.: Fused DNN: A deep neural network fusion approach to fast and robust pedestrian detection, *2017 IEEE Winter Conference on Applications of Computer Vision* (*WACV*), pp.953–961, IEEE (2017).

[7] Brazil, G., Yin, X. and Liu, X.: Illuminating Pedestrians via Simultaneous Detection & Segmentation, *Proc. IEEE International Conference on Computer Vision*, pp.1–10 (2017).

[8] Dollar, P., Wojek, C., Schiele, B. and Perona, P.: Pedestrian detection: A benchmark, *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp.304–311 (online), DOI: 10.1109/CVPR.2009.5206631 (2009).

[9] Dollar, P., Wojek, C., Schiele, B. and Perona, P.: Pedestrian detection: An evaluation of the state of the art, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.34, No.4, pp.743–761 (2011).

[10] Zhang, S., Benenson, R. and Schiele, B.: CityPersons: A diverse dataset for pedestrian detection, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.3213–3221 (2017).

[11] Neumann, L., Karg, M., Zhang, S., Scharfenberger, C., Piegert, E., Mistr, S., Prokofyeva, O., Thiel, R., Vedaldi, A., Zisserman, A., et al.: NightOwls: A pedestrians at night dataset, *Asian Conference on Computer Vision*, pp.691–705, Springer (2018).

[12] Viola, P. and Jones, M.: Rapid object detection using a boosted cascade of simple features, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.511–518 (2001).

[13] Tian, Y., Luo, P., Wang, X. and Tang, X.: Pedestrian detection aided by deep learning semantic tasks, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.5079–5087 (2015).

[14] Tian, Y., Luo, P., Wang, X. and Tang, X.: Deep learning strong parts for pedestrian detection, *Proc. IEEE International Conference on Computer Vision*, pp.1904–1912 (2015).

[15] Zhang, X., Cheng, L., Li, B. and Hu, H.: Too Far to See? Not Really!–Pedestrian Detection With Scale-Aware Localization Policy, *IEEE Trans. Image Processing*, Vol.27, No.8, pp.3703–3715 (online), DOI: 10.1109/TIP.2018.2818018 (2018).

[16] Brazil, G. and Liu, X.: Pedestrian Detection with Autoregressive Network Phases, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.7231–7240 (2019).

[17] Zhang, S., Yang, J. and Schiele, B.: Occluded pedestrian detection through guided attention in CNNs, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.6995–7003 (2018).

[18] Cai, Z., Fan, Q., Feris, R.S. and Vasconcelos, N.: A unified multi-scale deep convolutional neural network for fast object detection, *European Conference on Computer Vision*, pp.354–370, Springer (2016).

[19] Wang, S., Cheng, J., Liu, H. and Tang, M.: PCN: Part and context information for pedestrian detection with CNNs, arXiv preprint arXiv:1804.04483, pp.1–13 (2018).

[20] Li, J., Liang, X., Shen, S., Xu, T., Feng, J. and Yan, S.: Scale-aware fast R-CNN for pedestrian detection, *IEEE Trans. Multimedia*, Vol.20, No.4, pp.985–996 (2017).

[21] Du, X., El-Khamy, M., Morariu, V.I., Lee, J. and Davis, L.S.: Fused Deep Neural Networks for Efficient Pedestrian Detection, *CoRR*, Vol.abs/1805.08688, pp.1–11 (2018) (online), available from ⟨http://arxiv.org/abs/1805.08688⟩.

[22] Lin, C., Lu, J., Wang, G. and Zhou, J.: Graininess-aware deep feature learning for pedestrian detection, *Proc. European Conference on Computer Vision* (*ECCV*), pp.732–747 (2018).

[23] Song, T., Sun, L., Xie, D., Sun, H. and Pu, S.: Small-scale pedestrian detection based on somatic topology localization and temporal feature aggregation, arXiv preprint arXiv:1807.01438, pp.1–16 (2018).

[24] Nam, W., Dollár, P. and Han, J.H.: Local decorrelation for improved detection, arXiv preprint arXiv:1406.1134, pp.1–9 (2014).

[25] Dollár, P., Appel, R., Belongie, S. and Perona, P.: Fast feature pyramids for object detection, *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol.36, No.8, pp.1532–1545 (2014).

[26] Chang, C. and Lin, C.: LIBSVM: A library for support vector machines, *ACM Trans. Intelligent Systems and Technology*, Vol.2, No.3, pp.1–27 (2011).

[27] Uijlings, J., Sande, K., Gevers, T. and Smeulders, A.: Selective search for object recognition, *International Journal of Computer Vision*, Vol.104, No.2, pp.154–171 (2013).

[28] Ren, S., He, K., Girshick, R. and Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks, *Advances in Neural Informa-*

tion Processing Systems, pp.91–99 (2015).
[29] Appel, R., Fuchs, T., Dollar, P. and Perona, P.: Quickly boosting decision trees-pruning underachieving features early, *Proc. International Conference on Machine Learning*, pp.1–9 (2013).
[30] Noh, J., Lee, S., Kim, B. and Kim, G.: Improving occlusion and hard negative handling for single-stage pedestrian detectors, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.966–974 (2018).
[31] Redmon, J. and Farhadi, A.: YOLO9000: Better, faster, stronger, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.7263–7271 (2017).
[32] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y. and Berg, A.C.: SSD: Single shot multibox detector, *European Conference on Computer Vision*, pp.21–37, Springer (2016).
[33] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. and Fei-Fei, L.: Imagenet: A large-scale hierarchical image database, *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp.248–255, IEEE (2009).
[34] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A.: Going deeper with convolutions, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.1–9 (2015).
[35] Xie, S., Girshick, R., Dollár, P., Tu, Z. and He, K.: Aggregated residual transformations for deep neural networks, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.1492–1500 (2017).
[36] Chollet, F.: Xception: Deep learning with depthwise separable convolutions, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.1251–1258 (2017).
[37] Huang, G., Liu, Z., Van Der Maaten, L. and Weinberger, K.Q.: Densely connected convolutional networks, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.4700–4708 (2017).
[38] Yu, F., Koltun, V. and Funkhouser, T.: Dilated residual networks, *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pp.472–480 (2017).
[39] Li, Z., Peng, C., Yu, G., Zhang, X., Deng, Y. and Sun, J.: DetNet: A backbone network for object detection, arXiv preprint arXiv:1804.06215, pp.1–17 (2018).
[40] Brendel, W. and Bethge, M.: Approximating CNNs with bag-of-local-features models works surprisingly well on ImageNet, arXiv preprint arXiv:1904.00760, pp.1–15 (2019).
[41] Computing Receptive Fields of Convolutional Neural Networks, available from ⟨https://distill.pub/2019/computing-receptive-fields/⟩ (accessed 2020-01-17).
[42] 小平美沙季，王　彧，加藤ジェーン：マルチレベルの深層特徴を用いた歩行者検出，電子情報通信学会技術研究報告，pp.97–102 (2018).

王 彧

2013 年名古屋大学大学院情報科学研究科博士後期課程修了．2012 年 4 月～2016 年 3 月 JSPS 特別研究員．2016 年 4 月～2019 年 3 月名古屋大学特任助教，助教．現在，立命館大学情報理工学部助教．博士（工学）．

小平 美沙季

2018 年 3 月名古屋大学工学部卒業．2020 年 3 月同大学大学院情報学研究科博士前期課程修了．

加藤 ジェーン　（正会員）

1993 年名古屋大学大学院工学研究科情報工学専攻博士後期課程修了．同年富山大学工学部助手．1999～2000 年 Oxford 大学客員研究員．2000～2018 年名古屋大学大学院准教授．現在，立命館大学情報理工学部教授．移動物体追跡，物体詳細認識，行動認識，人物同定，映像要約，機械学習等に関する研究に従事．博士（工学）．電子情報通信学会等正会員．IEEE シニア会員．本会シニア会員．