Repulsion Lossを導入した Single Stage Headless構造による遠方歩行者検出

木村秋斗† 平川翼† 山下隆義† 藤吉弘亘†
†中部大学
E-mail: kim09@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，車が高速で運転している場合は，歩行者と車が近づく時間が短くなる．そのため，自動車周辺の歩行者検出だけでなく，遠方にいる歩行者を検出することは重要である．

従来の歩行者検出として，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]．しかし，遠方にいる歩行者は画像上では領域が小さく，従来の歩行者検出手法では，有効な特徴を得ることができない．また，従来の歩行者検出手法では，歩行者同士が重なっている場合に，複数の歩行者の特徴をまとめて一つにして検出することがある．これにより遠方の小さな物体の検出精度が低下する．一方でCNN を用いた顔検出手法は，遠方にいる小さい顔も含めた検出を行う手法が提案されている．

本研究は Single Stage Headless face detector(SSH)[4] を基に，Repulsion Loss[5] を損失関数に追加した手法を用いて遠方歩行者検出を行う．SSH は画像上の小さい顔を検出する手法である．そのため，SSH を歩行者検出に応用することで遠方にいる小さい歩行者の検出を行うことができる．また，Repulsion Loss は検出対象に対して予測された検出結果が他の検出対象から離れるように学習を行う損失関数である．これにより，歩行者同士が重なっている場合，歩行者同士をまとめて一つに検出することを防ぐことができる．本稿では，はじめに CityPersons データセットにより，SSH の遠方歩行者検出に最適な値の調査を行う．次に，遠方歩行者検出に最適な値に調整した SSH の損失関数に Repulsin Loss を追加し，追加前との精度の比較と従来の物体検出手法との比較を行い，遠方歩行者検出における本手法の有効性を調査する．

2 関連研究

歩行者検出手法はCNN を用いて歩行者の特徴を取得し，ネットワークから出力された歩行者かどうかの確率と歩行者領域の矩形座標から検出する手法が主流である．CNN を用いた歩行者検出手法は大きく two-stage 型と one-stage 型に分けられる．

2.1 two-stage 型

two-stage 型の歩行者検出手法は歩行者領域の推定を行った後，歩行者かどうかの分類を行う．中でも，Fast R-CNN[6] や Faster R-CNN[7] 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) により歩行者領域の推定を行う．RPN は，1:2，1:1，2:1 の 3 種類のアスペクト比ごとに大中小の 3 つのスケールを持った計 9 個の矩形のセット (アンカー) を事前に定義し，アンカーごとに歩行者かどうかのクラス確率と表示させる歩行者領域の座標を出力する．RPN により，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ことができるため，より高速な検出ができる．

2.2 one-stage 型

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，一枚の画像に対しての検出時間が速い．したがって，リアルタイムでの検出が求められる歩行者検出に応用される手法は one-stage 型の手法であることが多い．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格子状に分割し，格子内の領域ごとに，歩行者かどうかの分類と歩行者領域の推定を行うことで高速な歩行者検出を行うことができる．しかし，YOLO は分割する格子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．

SSD は畳み込み層ごとにクラス確率と歩行者領域の

座標を出力する．特徴マップはプーリングを行うごとに特徴マップが縮小する．そのため，入力に近いほど小さい物体を，出力に近いほど大きい物体の検出を行うことができるため，広範囲のスケールの歩行者検出を行うことができる．

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格子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．

3 Single Stage Headless face detector

SSH は SSD や YOLO と同様に one-stage 型の顔検出手法である．そのため，リアルタイムで顔の検出ができる．

SSH のネットワーク構造を図 1 に示す．SSH はベースネットワークの VGG-16[11] で取得した特徴をもとに 3 つのスケールごとの特徴マップを生成する．特に，小さい顔を検出するブランチではスケールの異なる特徴マップを融合し，より細かい特徴を取得する．これにより小さい顔も含めた広範囲のスケールの顔検出ができる．最後の検出モジュールでは，顔かどうかの分類と顔の位置推定を行う．

検出モジュールは図 2 に示すように 3×3 の畳み込み層とコンテキストモジュールにより詳細な特徴を取得し，1×1 の畳み込み層で顔かどうかの分類と位置の推定を行う．コンテキストモジュールは数層の 3×3 の畳み込み層により，歩行者のコンテキスト情報も含めた詳細な特徴の取得を行う．

位置推定は事前に定義したアンカーと正解位置とのオフセットを学習して行う．アンカーの定義は RPN と同様である．ただし，SSH はアンカーのアスペクト比を 1:1 のみに限定し，計算量の削減を行っている．

1 つの歩行者に対し 1 つの検出結果に限定するために Non-Maximum Suppression(NMS) を使用している．NMS は検出物体に対し複数のアンカーが閾値以上の Intersection over Union(IoU) を持つ場合，重なっているアンカーの中で最もクラス確率の高いアンカー以外のクラス確率を 0 にし，表示させないようにする．

SSH は検出モジュール-$k(k = 1, 2, 3)$ に対応するスケールのアンカーのクラス分類の損失と検出位置と正解位置とのオフセットを学習し，顔領域の特徴を取得する． SSH の損失関数を式 (1) に示す．

$$L_{SSH} = L_{Cls} + L_{Box} \tag{1}$$

ここで,L_{Cls} はクラス分類の損失，L_{Box} は検出位置と正解位置とのオフセットとする．L_{Cls}，L_{Box} をそれぞれ式 (2)，(3) に示す．

$$L_{Cls} = \sum_k \frac{1}{N_k^c} \sum_{i \in A_k} l_c(p_i, g_i) \tag{2}$$

$$L_{Box} = \sum_k \frac{1}{N_k^r} \sum_{i \in A_k} I(g_i = 1) l_r(b_i, t_i) \tag{3}$$

ここで，l_c を Cross Entoropy Loss，A_k を検出モジュール-k に対応するアンカー，p_i を i 番目のアンカーの予測カテゴリ，g_i を i 番目の正解ラベル,N_k^c を検出モジュール-k 内のスケールに対応するアンカー数，l_r を smooth L1 Loss，t_i をモジュールの i 番目のアンカーに対する正解位置， $I(\cdot)$ 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．

4 提案手法

本研究では，小さい顔を検出する手法である SSH のアンカーを調査し，遠方の歩行者検出に適したアンカーを用いる．また，小さい物体に発生しやすい検出対象同士の重なりを解消するために．損失関数に Repulsion Loss を導入し，より高精度な遠方歩行者検出を実現する．

4.1 アンカーの最適化

従来の SSH のアンカーサイズは 16 ピクセル，アスペクト比は 1:1 である．本研究では，SSH のネットワーク構造を使用し歩行者検出を行うため，アスペクト比を 1:2 の縦長のアンカーに変更し，歩行者検出に特化させる．また，歩行者検出の学習に用いる CityPersons データセット [12] の矩形の高さを調査したところ，図 3 に示すように，従来のアンカーサイズである 16 ピクセル以下の矩形が多数存在していることを確認した．現状では遠方にいる小さい歩行者の特徴を学習することができないため，アンカーサイズを小さくし，遠方歩行者の特徴を取得する必要がある．しかし，アンカーサイズを過度に縮小すると，極小の特徴から歩行者を推定するため誤検出が発生する．そこで，アンカーサイズを変更した場合の距離ごとの精度比較による調査から遠方歩行者検出に適したアンカーサイズを決定する．

4.2 Repulsion Loss

歩行者同士が重なっているとき，ネットワークが複数の歩行者をまとめて一つの歩行者の特徴として取得する場合がある．その結果，複数の歩行者の領域をまとめて検出することで，NMS により他の検出結果を削除する問題がある．Repulsion Loss[5] は，異なる検出対象の矩形の位置を離すように学習する損失関数である．そのため，複数の歩行者が重なっていた場合，重なりが少なくなるように学習され，NMS による検出結果の削除を防ぐことができる．Repulsion Loss を式 (4) に示す．

$$L_{Rep} = L_{Box} + \alpha * L_{RepGT} + \beta * L_{RepBox} \tag{4}$$

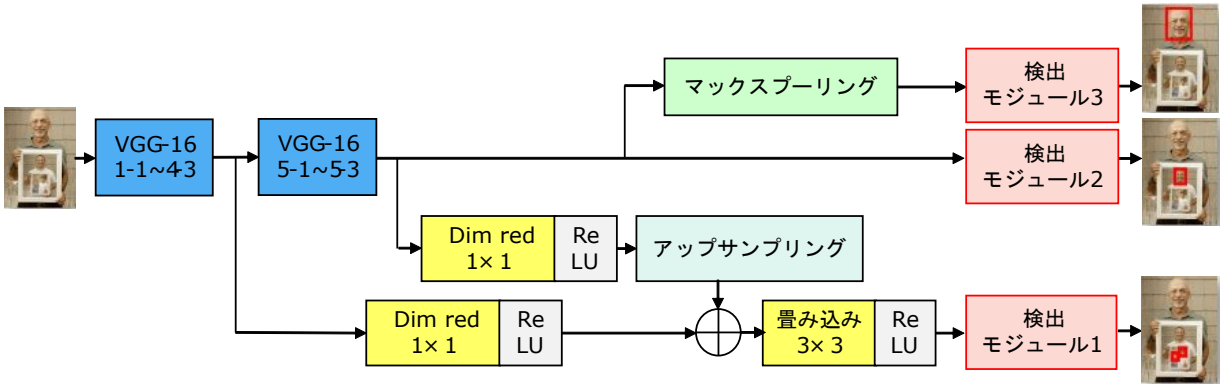


図 1　SSH の構造

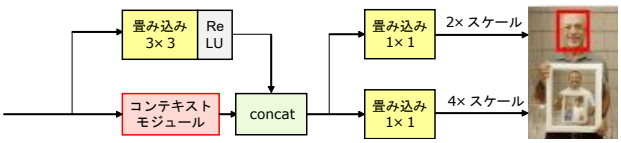


図 2　検出モジュールの構造

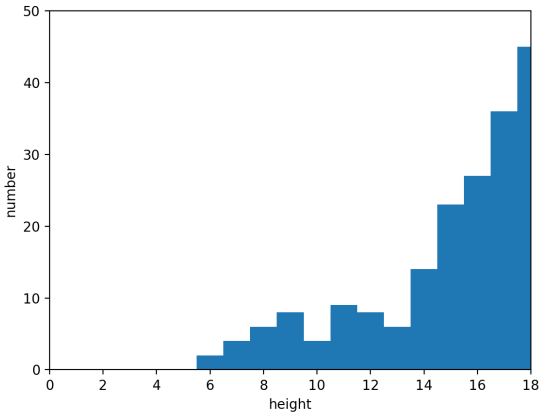


図 3　学習用データセットの小矩形分布

ここで α，β は重みである．L_{Box} は SSH の検出位置と正解位置とのオフセットである．L_{Box} を式 (5) に示す．

$$L_{Box} = \frac{\sum_{i \in P} l_1(b_i, g_i)}{P} \tag{5}$$

ここで，b_i は検出位置，P は正解と定義されたアンカー数である．L_{RepGT} は予測した矩形に対して異なる検出対象の正解歩行者位置から遠ざけるように学習する．L_{RepGT} を式 (6) に示す．

$$L_{RepGT} = \frac{\sum_{i \in P} l_n(\frac{b_i \cap g_{Rep}}{g_{Rep}})}{P} \tag{6}$$

ここで，l_n は smooth Ln Loss である．g_{Rep} は検出位置との IoU が最も高い異なる検出対象の正解位置を使用する．smooth Ln Loss はハイパーパラメータである σ の値によって，傾きが変化する．L_{RepBox} を式 (7) に示す．L_{RepBox} は予測された矩形に対して，異なる検出対象の検出位置から遠ざけるように学習する．

$$L_{RepBox} = \frac{\sum_{i \neq j} l_n(b_i, b_j)}{\sum_{i \neq j} \mathbb{I}[IoU(b_i, b_j) > 0] + \epsilon} \tag{7}$$

b_j は他の検出対象の検出位置，$\mathbb{I}(\cdot)$ は指定した値をそのまま返す恒等関数，ϵ は分母を 0 にしないための極小の値である．検出位置と異なる検出対象の重なりがない場合，Repulsion Loss は 0 となる．

4.3　Repulsion Loss の導入

SSH の学習の際に，アンカーの回帰損失として，Repulsion Loss を追加する．Repulsion Loss を追加した SSH 全体の損失関数 L_{SSH} を式 (8) に示す．

$$L_{SSH} = L_{Cls} + L_{Box} + \alpha * L_{RepGT} + \beta * L_{RepBox} \tag{8}$$

L_{Box} は SSH のアンカーの回帰損失を使用するため，各検出モジュール-k に対応するスケールのアンカーに限定して損失が与えられる．それに伴い，本研究の Repulsion Loss は式 (9)，(10) に示すように各検出モジュール-k に対応するスケールの検出位置に発生する検出対象との重なりのみを考慮する．

$$L_{RepGT} = \sum_k \frac{\sum_{i \in P_k} l_n(\frac{b_i \cap g_{Rep}}{g_{Rep}})}{P_k} \tag{9}$$

$$L_{RepBox} = \sum_k \frac{\sum_{i \neq j \in P_k} l_n(b_i, b_j)}{\sum_{i \neq j \in P_k} \mathbb{I}[IoU(b_i, b_j) > 0] + \epsilon} \tag{10}$$

本研究の smooth ln Loss の σ は，[5] を参考に $\sigma = 0.5$ とする．

5　評価実験

本実験では遠方歩行者の検出に適したアンカーサイズの調査を行う．また，従来の SSH の損失関数と Repulsion Loss を追加した損失関数を用いた場合の精度を比較し，Repulsion Loss の有効性を調査する．

5.1　データセット

本実験の学習，評価には CityPersons データセットを使用する．CityPersons データセットはヨーロッパの都市で撮影された車載画像データから構成されている．本実験では学習画像に 2303 枚，評価画像に 398 枚を用いる．入力画像のサイズは，遠方の小さい歩行者の消失を防ぐために画像を縮小せず，2048×1024 ピクセルとする．

5.2　実験概要

アンカーサイズの調査では，アンカーのサイズが 4，8，12，16 ピクセルの時の精度を比較する．評価は Cityscapes データセット [13] のデプス画像をもとに算出した距離別に行う．距離 $distance$ の算出式を式 (11) に示す．

$$distance = \frac{baseline * f}{disp} \tag{11}$$

ここで，$baseline$ をカメラ間の視差，f をピクセル単位の焦点距離，$disp$ をデプス画像の画素値とする．

また，定量的な比較に mean Average Precision(mAP) を使用する．mAP を式 (12) に示す．

$$mAP = \frac{1}{M}\frac{1}{TP+FP}\sum_{i\in(TP+FP)} precision_i \tag{12}$$

ここで，M を評価画像枚数，TP を正解と判断した検出結果数，FP を不正解と判断された検出結果数とする．また，$precision$ を式 (13) に示す．

$$precision = \frac{TP}{TP+FP} \tag{13}$$

本実験では検出結果と正解との IoU が 0.5 以上のものを正解と定義する．

Repulsion Loss の有効性の調査でのアンカーのサイズは，アンカーサイズの調査実験で最も精度の良いものを使用する．定量的な評価は mAP，$precision$，$recall$ の三種類を使用する．$recall$ を式 (14) に示す．

$$recall = \frac{TP}{TP+NP} \tag{14}$$

ここで，NP を正解位置のうち未検出であったものの数とする．全体の検出結果の評価と合わせて，距離が 80m 以上離れているものを遠方歩行者と定義し，比較を行う．また，提案手法と CityPersons データセットにより学習，評価を行った YOLOv3 との比較を行い，本手法の有効性を調査する．

学習は，バッチサイズ 1 で学習回数は 45000 iteration，最適化関数に momentum SGD を使用する．momentum は 0.9，重みの減衰は 0.0005，学習率は学習初期は 0.004，18000iteration 終了後は 0.0004 に設定する．

5.3　アンカー最適化

各アンカーサイズにおける距離ごとの精度を図 4 に示す．80m 以上の検出において最も精度が高かったアンカーサイズは 8 ピクセルであることがわかる．80m より手前の精度はアンカーサイズが小さいほど精度が悪くなった．これは小さいアンカーサイズでは，手前にいるサイズの大きな歩行者領域を捉えきることができないからである．

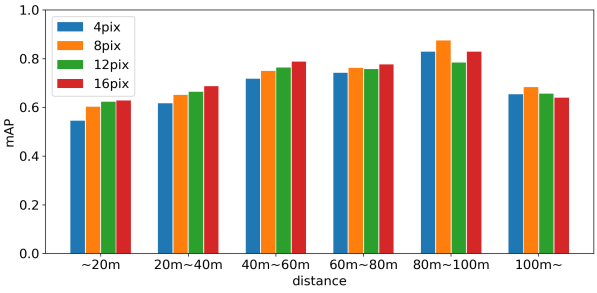


図 4　各アンカーサイズにおける mAP

5.4　Repulsion Loss の導入

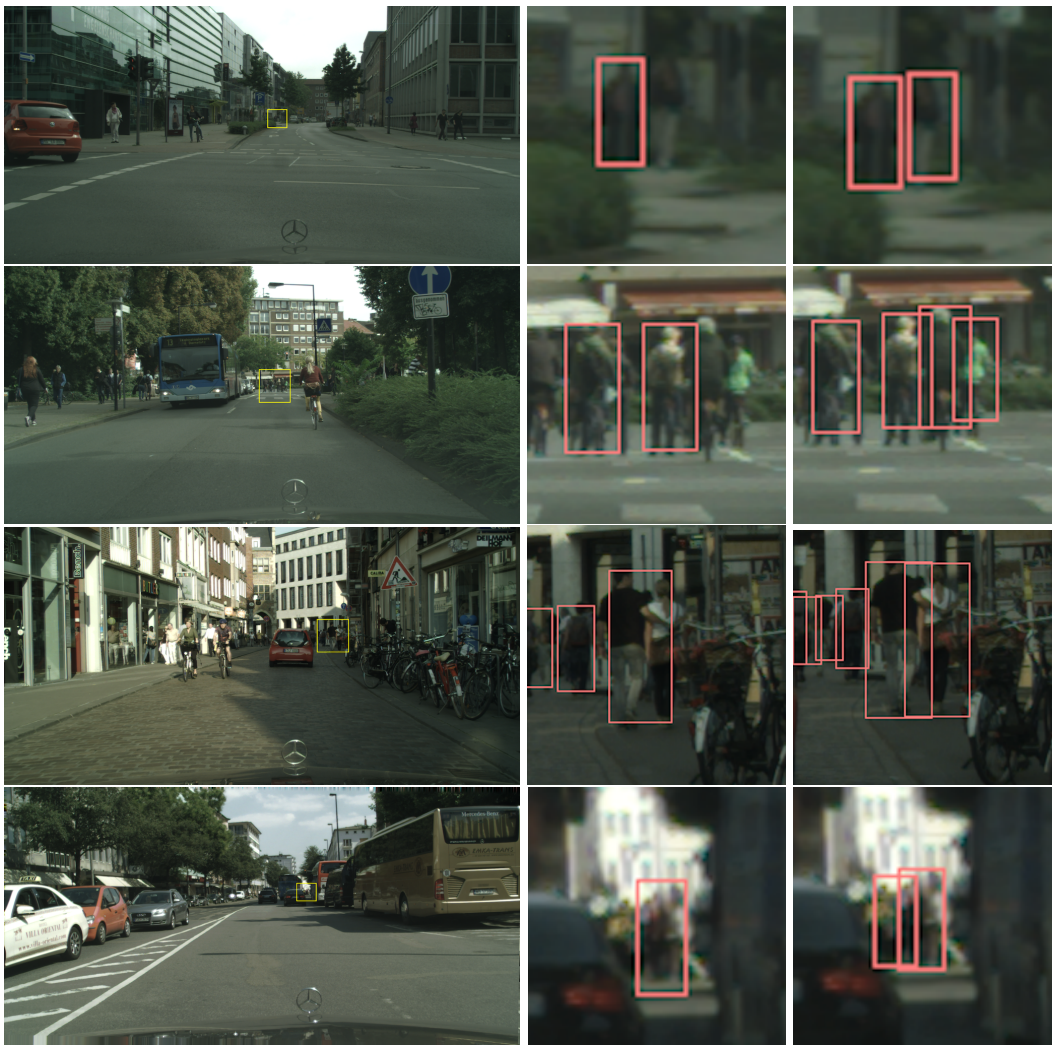
従来の SSH による検出精度と Repulsion Loss を追加した検出精度の結果を表 1 に示す．両手法のベースアンカーサイズはアンカーサイズの調査結果から 8 ピクセルとしている．これより，遠方歩行者の検出において mAP が 4.6 ポイント，recall が 3.9 ポイント増加したことから遠方歩行者の未検出が減少したことにより精度が良くなることがわかる．YOLOv3 との比較結果を表 2 に示す．遠方の歩行者検出において mAP が 9.6 ポイント，recall が 15.9 ポイント増加したことから，提案手法が従来の物体検出手法による歩行者検出よりも適していることがわかる．また，提案手法の検出例を図 5 に示す．図 5 から Repulsion Loss の導入により，重なりによって未検出であった遠方歩行者が検出できていることがわかる．

表 1　損失関数変更による歩行者の検出精度

	mAP	precision	recall
従来手法	59.1	77.1	61.3
提案手法	**64.1**	67.2	**66.9**
従来手法 (遠方)	64.8	79.7	65.9
提案手法 (遠方)	**69.4**	77.9	**69.8**

表 2　他手法との精度比較

	mAP	precision	recall
YOLOv3	49.1	80.2	50.4
提案手法	**64.1**	67.2	**66.9**
YOLOv3(遠方)	59.8	82.0	53.9
提案手法 (遠方)	**69.4**	77.9	**69.8**



(a) 元画像　(b) 従来手法　(c) 提案手法

図 5　SSH による歩行者検出結果

6　おわりに

本研究では，小さい顔を検出する手法である SSH の遠方歩行者検出に最適なアンカーサイズの調査を行った．また，歩行者同士の重なりに対応するための損失関数である Replusion Loss を導入した．

SSH の CityPersons データセットにおけるアンカーの最適化により，遠方歩行者に最適なアンカーサイズが 8 ピクセルであることを確認した．また，Repulsion Loss の導入により，導入前と比較して，遠方歩行者の検出において mAP が 4.6 ポイント，recall が 3.9 ポイント向上し，YOLOv3 と比較して mAP が 9.6 ポイント，recall が 15.9 ポイント上回った．このことから遠方歩行者検出用に特化させた SSH に Repulsion Loss を導入することでより高精度な遠方歩行者の検出ができることを確認した．

今後は，歩行者以外の誤検出を抑制しながらも，より高速な検出を行う手法を考案する．

参考文献

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

[2] Bin Yang, Junjie Yan, Zhen Lei, and Stan Z. Li. Convolutional channel features. In *The IEEE International Conference on Computer Vision (ICCV)*, pages 82–90, December 2015.

[3] Jianan Li, Xiaodan Liang, Shengmei Shen, Tingfa Xu, and Shuicheng Yan. Scale-aware fast r-cnn for pedestrian detection. In *IEEE Transactions on Multimedia*, pages 985–996, 2017.

[4] Najibi Mahyar, Samangouei Pouya, Chellappa Rama, and Davis Larry, S. Ssh: Single stage headless face detector. pages 4875–4884, 2017.

[5] Wang Xinlong, Xiao Tete, Jiang Yuning, Shao Shuai, Sun Jian, and Shen Chunhua. Repulsion

loss: Detecting pedestrians in a crowd. In *the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7774–7783, 2018.

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

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

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

[9] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. 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] Russakovsky Olga, Deng Jia, Su Hao, Krause Jonathan, Satheesh Sanjeev, Ma Sean, Huang Zhiheng, Karpathy Andrej, Khosla Aditya, Bernstein Michael, Berg Alexander, C, and Fei-Fei Li. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[12] Zhang Shanshan, Benenson Rodrigo, and Schiele Bernt. Citypersons: A diverse dataset for pedestrian detection. In *the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017.

[13] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3223, 2016.