Pixel-wise-Attention による Point 型歩行者検出の判断根拠の可視化

木村　秋斗[1*)]　長内　淳樹[2)]　平川　翼[1)]　山下　隆義[1)]　藤吉　弘亘[1)]

Visualizing the basis of Point-base pedestrian detection using Pixel-wise-Attention

Shuto Kimura　Atsuki Osanai　Tsubasa Hirakawa　Takayoshi Yamashita　Hironobu Fujiyoshi

Pedestrian detection is a necessary technology to ensure safety in automated driving. In order to realize a safer and more secure automated driving, it is necessary to clarify the basis of the pedestrian detection. Conventional pedestrian detection methods use Convolutional Neural Network (CNN) object detection to detect pedestrians. Several methods use Region Proposal Network (RPN), which predicts the position of a pedestrian by determining the offset between the pedestrian and a set of rectangular boxes called anchors. RPNs compute coordinates and size offsets, there are other areas to focus on besides pedestrians. However, they are not clear as a basis of judgment for pedestrian detection. In this study, we visualize the basis of judgment for pedestrians by using Pixel-wise-Attention in the Point-base object detection method, which detects without anchors. This method introduces a Pixel-wise-Attention mechanism to obtain the attention map for an arbitrary pixel in the Center and Scale Prediction (CSP), which detects pedestrians based on the center of the detection target and the scale from that point. By acquiring the attention map of the center of the estimated pedestrian, we can obtain pedestrians and the surrounding area of interest as a basis for judgment. In our experiments, we visualize the attention area of the CSP with the Pixel-wise-Attention mechanism and show the basis for the decision of pedestrian detection.

KEY WORDS: **safety, accident avoidance, image processing,** pedestrian detection, attention

1. まえがき

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．自動運転における歩行者検出は，車の前方にいる歩行者を検出し，車を静止させることで，安全を確保するための必要な技術である．また，より安心で安全な歩行者検出を実現するためには，歩行者検出は何を根拠に判断したかを明確にする必要がある．

従来の歩行者検出の手法として，Convolutional Newral Network(CNN)(1)による物体検出を利用し，歩行者を検出する手法が提案されている(2〜6)．中でも，歩行者の位置推定には，アンカーと呼ばれる矩形の雛形のセットと歩行者のずれを求めることで歩行者の位置を検出する Region Proposal Network(RPN)(4)を用いる手法が多く提案されている．しかし，RPN は座標や大きさのずれを求める性質上，歩行者そのものの情報以外にも注目すべき箇所があり，歩行者検出の判断根拠としては不明確であるといえる．

本研究は，アンカーを必要としない Point 型物体検出手法に Pixel-wise-Attention を用いることで，歩行者の判断根拠の可視化を行う．本手法は，検出対象の中心とその地点からのスケールを基に，歩行者を検出する Center and Scale Prediction(CSP)(7)に任意の画素に対する注視領域を取得する Pixel-wise-Attention 機構を導入する．これにより，推定した歩行者の中心に対する注視領域の獲得で，歩行者およびその周辺の注目領域を判断根拠として獲得できる．本稿では，Pixel-wise-Attention 機構を導入した CSP の注視領域を可視化し，歩行者検出の判断根拠を示す．また，特徴に対し注視領域を重み付けしたものと従来の CSP で精度の比較を行い，注視領域を考慮した歩行者検出の有効性を調査する．

1)　中部大学(487-8501 愛知県春日井市松本町 1200)
2)　(株)本田技術研究所(461-2511 埼玉県和光市中央１丁目４番１号)
*) 講演者

2. 関連研究

2.1. 歩行者検出手法

歩行者検出手法は CNN を用いて歩行者の特徴を取得し，ネットワークから歩行者かどうかの確率と推定した歩行者の位置を出力して検出する手法が主流である．代表的な位置推定の方法にアンカーを用いる手法と対象の Point を検出する手法の 2 つがある．

2.1.1. アンカーを用いた検出手法

歩行者検出手法の多くは RPN により歩行者の検出を行う．RPN はアンカーと呼ばれる大きさとアスペクト比が異なる矩形の雛形を事前に定義し，アンカー1 つ 1 つに対して歩行者かどうかの確率と正解である歩行者領域から座標および大きさがどの程度ずれているのかを出力する．RPN を用いることで，単一のネットワークで歩行者かどうかの分類と歩行者の位置推定が同時にできるようになり，高速かつ高精度な検出を実

現する．しかし，RPN による位置推定は事前定義したアンカーの構造に依存するため，極端な大きさやアスペクト比の検出が困難である．

2.1.2. Point 型検出手法

Point 型物体検出手法は，対象の歩行者の中心や上下端などを Point と定義し，Point の画素を基に歩行者検出を行う．Point 型物体検出はアンカーの事前定義を必要としないため，大きさやアスペクト比に依存しない検出ができる．Point 型物体検出手法は対象物体の様々な箇所を Point とした手法が提案されている(7～10)．Center-and-Scale Prediction(CSP)は歩行者の中心点を Point と定義し，Point とそこからの歩行者のスケールを基に歩行者を検出する．CSP は歩行者の中心かどうかの確率，中心からの歩行者のスケール，画像を元の大きさに戻した際にどの程度ずれるのかの 3 つのパラメータで歩行者を検出する．そのため，事前定義したアンカーで推定を行う RPN に対し，推定のコストを大幅に削減することができる．

2.2. 注視領域の可視化

深層学習ネットワークの判断根拠を示すための注視領域は，ネットワークから取得した特徴から空間方向の attention を取得することで生成される attention map を可視化することで求めることができる．Liu ら(11)は Local Attention Pooling 構造により，対象の画素に対する attention map を取得する PiCANet を提案している．Local Attention Pooling 構造ではネットワークから取得した特徴マップから任意の範囲を切り抜き，畳み込み処理によって周囲の情報と範囲内の総画素数分の attention map を取得する．これにより，範囲内の任意の画素に対する attention map を取得することができる．

3. Attention-Mask 機構用いた判断根拠の可視化

本章は Attention-Mask 機構を導入した物体検出ネットワークの判断根拠の可視化を行う．Attention-Mask 機構によって生成された attention map を可視化することで判断根拠の調査を行う．

3.1. Attention-Mask 機構

Attention-Mask 機構を図 1 に示す．ベースネットワークで取得した特徴を基に画像内の歩行者領域に注視した Grid-Attention Map(GAM)を生成する．生成した GAM を用いて特徴マップに重み付けすることで注視領域を考慮した歩行者検出ができる．GAM は画像内の歩行者教師データの歩行者領域内を 1，それ以外を 0 とした教師 GAM による学習を行う．GAM の損失式を式(1)に示す．

$$L_{GAM} = \mathrm{MSE}(A_p, A_t) + \sum_{k=0}^{K} \mathrm{MSE}(A_{p_k}, A_{t_k}) \qquad (1)$$

ここで，A_pは Attention-Mask 機構で生成された GAM，A_tは教師データを基に生成された教師GAM，A_{p_k}, A_{t_k}はそれぞれA_p，

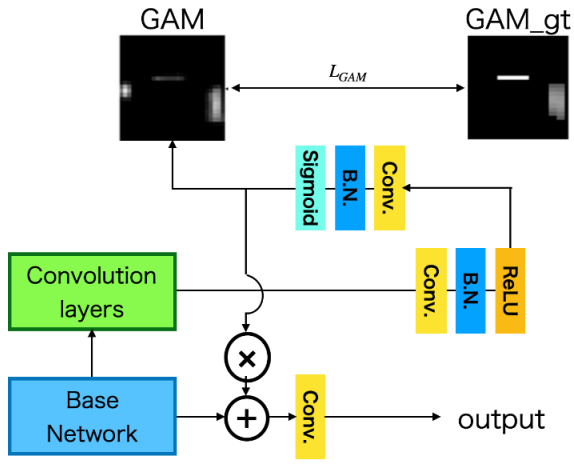


Fig.1 Attention-Mask mechanism

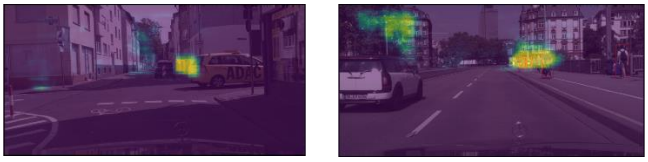


Fig.2 Visualization of GAM

A_tを正解矩形の範囲内に限定したもの，Kは画像内の正解矩形の数である．

3.2. Attention-Mask 機構による判断根拠

Attention-Mask 機構によって生成された GAM の可視化結果を図 2 に示す．生成された GAM は歩行者領域全体を注視していることがわかる．Attention-Mask 機構によって生成された GAM は歩行者領域を囲う矩形を教師データとして学習している．そのため，Attention-Mask 機構における注視領域は歩行者およびその周辺ではなく，歩行者の領域を囲う矩形に対して注視されたものであると考えられる．

4. 提案手法

本研究では，Point 型検出手法である CSP に Pixel-wise-Attention 機構を導入し，画素に対する attention map を取得することで歩行者およびその周辺を注目した歩行者検出ネットワークの判断根拠の可視化を行う．

4.1. ネットワーク構造

本手法の構造を図 3 に示す．任意の入力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．ベースネットワークで取得した特徴マップを基にスケールの異なる 3 つの特徴マップを得ることで様々な大きさの歩行者の検出ができるようになる．3 つの特徴マップをそれぞれスケールの大きいものから順に Scale1，Scale2，Scale3 の Pixel-wise-Attention 機構に入力し，各画素に対しての attention

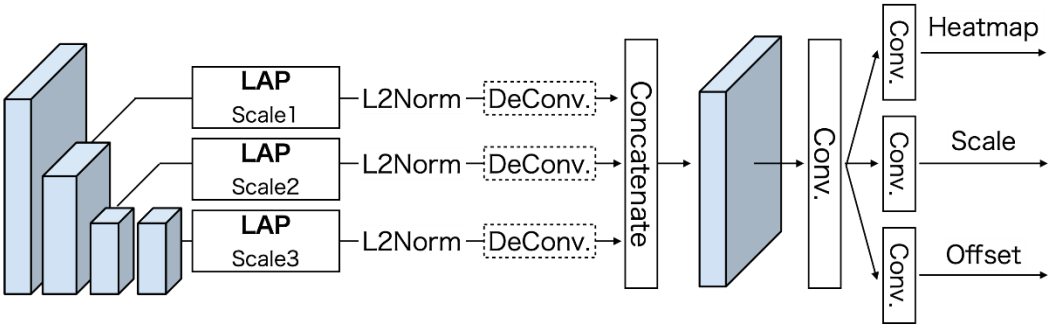


Fig.3 Structure of proposed method

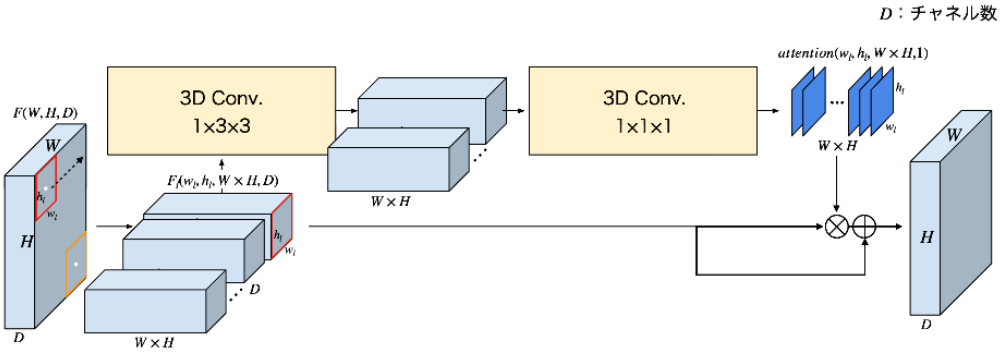


Fig.4 Pixel-wise-Attention mechanism

map を取得する.そして,入力した特徴マップに attention map を重み付けする. 重み付けされた 3 つの特徴マップを合わせるために Deconvolution 層を適用する.3 つの特徴マップを合統合した後, 各画素に対して歩行者の中心かどうかの確率を示す Heatmap, 中心点からの大きさを示す Scale, 元の画像の大きさに戻した際に発生する中心点のずれを示す Offset を各 Convolution 層によって出力する.

4.2. Pixel-wise-Attention 機構

Pixel-wise-Attention 機構を図 4 に示す. Pixel-wise-Attention 機構では大きさ$h_l \times w_l$の attention map を入力特徴マップの総画素数分取得し, 重み付けをする.

入力した特徴マップ$\mathrm{F} \in \mathbb{R}^{\mathrm{H}\times\mathrm{W}\times\mathrm{D}}$を左上端の画素から対象画素を中心とした周囲$h_l \times w_l$を sliding window で順にスキャンし, 入力特徴マップの総画素数である$\mathrm{W} \times \mathrm{H}$個分のデータ$F_l \in \mathbb{R}^{h_l \times w_l \times (\mathrm{W}\times\mathrm{H})\times\mathrm{D}}$を系列データとして取得する. そして, 系列データを空間方向に畳み込み, 周囲の情報を取得する. その後, チャネル方向に畳み込みを行い, $\mathrm{W} \times \mathrm{H}$個分の attention map を取得する. 本手法では計算コスト削減のため, 系列データに対する畳み込みには系列方向に対して重みを共有する 3 次元畳み込みを使用する. 作成した attention map を系列データF_lに対して重み付けをしたものを出力する. これにより, 注視領域を考慮した歩行者検出ができる. 重み付け時, 特徴が極端になることを防ぐために残差機構を導入する.

4.3. 学習方法

本手法は CSP と同様に出力した Heatmap, Offset を基に推定した歩行者の中心, Scale を基に推定した歩行者のスケールと正解との差異を損失とし, 少なくするように学習を行う. 正解は学習を容易にするために正解座標の周囲に式(2)に示すガウシアンマスクM_{ij}による情報を付与する.

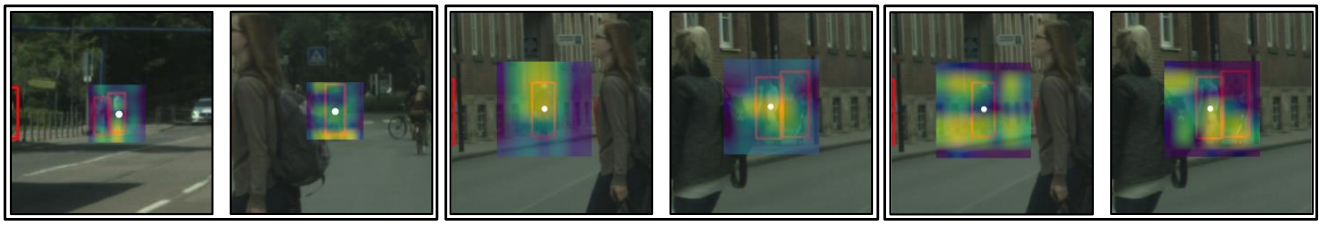
$$M_{ij} = \max_{k=1,2\ldots K} G(i,j;x,y,\sigma_w,\sigma_h)$$
$$G(i,j;x,y,\sigma_w,\sigma_h) = e^{-\left(\frac{(i-x)^2}{2\sigma_w^2}+\frac{(i-y)^2}{2\sigma_h^2}\right)} \quad (2)$$

ここで, i, jはそれぞれ対象位置の座標, x, yは正解歩行者の中心位置の座標, σ_w, σ_hはそれぞれ中心位置からの幅と高さ, Kは画像内に存在する歩行者の数である. マスクのより正解歩行者の周囲にも中心位置の情報が与えられ, スケール推定した歩行者の中心からわずかにずれた場合でも歩行者の中心位置の学習ができるため, 効率的な学習を行うことができる.

歩行者の中心の損失L_{center}を式(3)に示す.

$$L_{center} = -\frac{1}{K}\sum_{i=1}^{W}\sum_{j=1}^{H} a_{ij}(1-\hat{p}_{ij})^2 \log(\hat{p}_{ij}) \quad (3)$$

ここで, W, Hは合成した特徴マップの幅と高さ, α_{ij}はガウシアンマスクによる重みであり, $\hat{p}_{ij}$は推定した座標ごとの歩行者の中心かどうかの確率p_{ij}に基づいた値である. α_{ij}は対象座

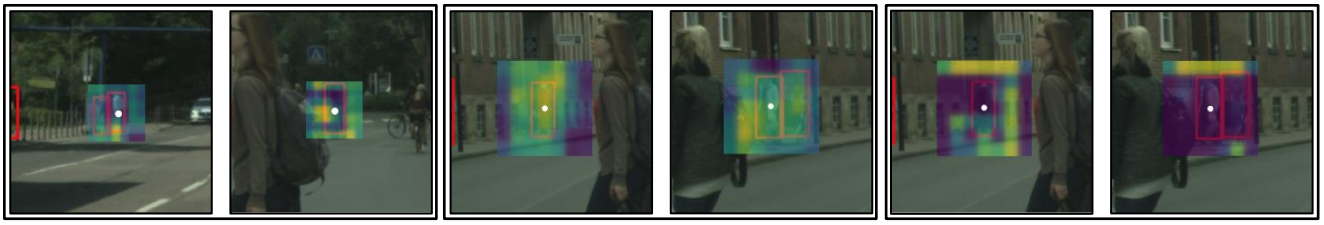


(a) Scale1　　(b) Scale2　　(c) Scale3

Fig.5 Visualization of attention map

(a) Scale1　　(b) Scale2　　(c) Scale3

Fig.6 Visualization of attention map in case that the network is trained only with center



(a) Scale1　　(b) Scale2　　(c) Scale3

Fig.7 Visualization of attention map in case that the network is trained only with scale

標が正解歩行者の中心位置の場合は，$\alpha_{ij} = 1$，それ以外は$\alpha_{ij} = (1 - M_{ij})^4$を与える．$\hat{p}_{ij}$は対象座標が正解歩行者の中心位置の場合は，$\hat{p}_{ij} = p_{ij}$，それ以外は$\hat{p}_{ij} = 1 - p_{ij}$を与える．スケールの損失$L_{scale}$を式(4)に示す．

$$L_{scale} = \frac{1}{K}\sum_{k=1}^{K} SmoothL1(s_k, t_k) \tag{4}$$

ここで，s_kは推定されたスケール，t_kは正解のスケールである．

5. 評価実験

本実験では，提案手法によって生成された attention map を可視化し，Point 型歩行者検出ネットワークの判断根拠を示す．また，提案手法の精度による比較実験から，注視領域を考慮した歩行者検出の有効性を調査する．

5.1. 実験概要

本実験の学習，評価には CityPersons データセット(12)を使用する．CityPersons データセットはヨーロッパ市街で撮影された車載画像および画像内の歩行者の位置情報のデータで構成されている．本実験では学習用に 2975 枚，評価用に 500 枚の画像を使用する．学習は 100 epoch 行う．

判断根拠の調査では検出された歩行者の中心位置の画素に対しての attention map を可視化する．また，CSP の学習対象を中心点のみ，スケールのみに限定した場合の attention map を可視化し，注視領域の比較を行う．

比較実験では従来の CSP と本手法の比較を行い，注視領域を考慮した歩行者検出が精度向上に有効であるかを調査する．また，Pixel-wise-Attention 機構の系列データ作成時，特徴マップからランダムに値を取得することで作成した系列データと sliding window によって値を取得し作成した系列データによる精度検証を行い，Pixel-wise-Attention 機構がベースネットワークから取得した特徴に基づいた注視領域の獲得と，重み付けができているかを確認する．さらに，Attention-Mask

機構を導入した CSP 本手法の比較を行い，歩行者およびその周辺を注視した注視領域による重み付けが有効であるか調査する．定量的な評価指標として log-average miss rate(13)を使用する．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)の平均をから求める．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，50 pixel 以上を Middle，50 pixel 未満のものを Small とし，大きさごとの比較を行う．

5.2. 判断根拠の可視化

Pixel-wise-Attention 機構によって取得された attention map を図 5 に示す．図の attention map は白点の画素に対しての attention map であり，黄色に近いほど強く反応していることを示す． attention map は歩行者の頭部付近や足元を強く注視している．また，出力した attention map の大きさが検出対象の大きさに近い場合は，特徴マップのスケールの違いにかかわらず歩行者の中心から上下端にかけて強く注視する傾向がみられる．このことから，CSP は対象画素が歩行者の中心である場合，頭部や足元を判断根拠としていることがわかる．

学習対象を中心点のみにした場合の attention map を図 6 に，スケールのみにした場合の attention map を図 7 に示す．中心のみを学習した attention map には全体を強く注視する傾向が見られる．このことから，対象画素が歩行者の中心かどうかは，歩行者の周囲の情報から判断していることがわかる．スケールのみを学習した attention map では Scale1 で歩行者の下部分を，Scale3 では歩行者の上部分を強く注視し，Scale2 では歩行者全体に広く注視する傾向が見られる．このことから，スケールの推定には歩行者の上端や下端を基準に判断していることがわかる．本手法で生成される attention map が違いはあるものの，スケールごとに傾向が大きく変化はしない．これは，本手法の構造上各スケールの特徴を合わせてから推定を行うため，スケールごとの注視領域も他のスケールと合わせるように学習しているためである．

5.3. 検出精度の比較

従来の CSP と提案手法の比較結果を表 1 に示す．提案手法は従来の CSP と比較し Large で 1.9 ポイント，全体で 1.1 ポイント低下している．このことから，歩行者検出において注視領域を考慮することは有効であることがわかる．

Table1 Log-average miss rate over different scale

	All	Large	Middle	Small
CSP	15.8	23.2	**5.8**	**7.3**
proposed method	**14.7**	**21.3**	6.8	7.8

また，系列データ作成の際特徴マップからランダムに取得した系列データを用いる場合と，sliding window で取得した系列データによる Pixel-wise-Attention 機構を導入した場合の比較結果を表 2 に示す．ランダムに取得した系列データでは log-average miss rate が 100%に近くほとんど検出ができていない．このことから，本手法はベースネットワークから取得した特徴に基づいた注視領域の獲得と重み付けができていることがわかる．

Table2 Comparison of the series data acquisition

	All	Large	Middle	Small
ランダム値	98.8	97.6	99.1	100
proposed method	14.7	21.3	6.8	7.8

Attention-Mask 機構を導入した CSP と本手法の比較結果を表 3 に示す．提案手法は Attention-Mask 機構を導入した CSP に比べ， Middle で 1.4 ポイント，Small で 3.4 ポイント，全体で 5.0 ポイント低下している．このことから，歩行者およびその周辺の注視領域による重み付けを行うことで，より高精度な歩行者検出ができる．

Table3 Comparison with Attention-Mask mechanism

	All	Large	Middle	Small
Attention-Mask	19.7	**20.3**	8.2	11.2
proposed method	**14.7**	21.3	**6.8**	**7.8**

6. まとめ

本研究は，歩行者の任意の箇所の画素を基に検出する Point 型検出手法に画素に対しての Pixel-wise-Attention 機構を導入することで歩行者検出の判断根拠の可視化を行った．出力された attention map から CSP では歩行者の頭部や足元を強く注視する傾向が見られた．また，従来の CSP との比較実験では導入前と比較し，Large で 1.9 ポイント，全体で 1.1 ポイントの精度向上が見られた．このことから，注視領域を考慮することでより高度な歩行者検出が実現できることを確認した．

今後は，より高解像度な attention map を出力し，より詳細な判断根拠の可視化を目指す．

参 考 文 献

(1) Yann LeCun, Leon 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, 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) 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.
(5) 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.
(6) Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. volume abs/1804.02767, 2018.
(7) Wei Liu, Shengcai Liao, Wiqiang Ren, Weidong Hu 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.
(8) Hei Law and jia Deng. CornerNet:Detecting Objects as Paired Keypoints. In European Conference on Computer Vision, pages 734-750, 2018.
(9) Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang and Qi Tian. CenterNet: Keypoint Triplets for Object Detection. In the IEEE International Conference on Computer Vision, pages 6569-6578, 2019.
(10) Tao Song, Leiyu Sun, Di Xie, Haiming Sun and Shiliang Pu. Small-scale pedestrian detection based on topological line localization and temporal feature aggregation. In: The European Conference on Computer Vision, pages 536-551, 2018.
(11) Mian Liu, Junwei Han and Ming-Hsuan Yang. PiCANet:Lieraning Pixel-wise Contextual Attention for Saliency Detection. In the IEEE Conference on Computer Vision and Pattern Recognition, pages 3089-3098, 2018.
(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) Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. IEEE transactions on pattern analysis and machine intelligence 34(4), pages 743-761, 2012.