遠距離物体検出に適したネットワークアーキテクチャ探索手法

平川 翼[1)]　　山下 隆義[1)]　　藤吉 弘亘[1)]

Architecture Search for Distant Object Detection

Tsubasa Hirakawa　　Takayoshi Yamashita　　Hironobu Fujiyoshi

Object detection is an important task in autonomous driving. Especially, detecting not only near but also distant objects in a driving scene is crucial for safety driving and further development of autonomous driving applications such as traffic light recognition and path predictions of pedestrians. While deep learning-based object detection methods now becomes a common approach, the architecture of these methods are manually designed and developed. In this paper, we propose a method to find an optimal network architecture for distant object detections. Our method is based on a neural architecture search (NAS), and the method searches an optimal architecture during training automatically. The experimental results with onboard vehicle camera image dataset for pedestrian detection show that our found optimal network architecture successfully detect distant objects.

KEY WORDS: **Electronics and Control, Image Recognition System, Image Processing,** Object Detection (E1)

1. ま え が き

自動運転において，周囲の環境を把握することは適切な車両制御の実現や交通事故を未然に防ぐための重要な要素である．周囲環境を把握するためには，車載カメラや Light Detection and Ranging (LiDAR) 等を用いて環境をセンシングし，センシングデータを用いて認識を行うことで実現される．認識には，車両周囲の歩行者や自動車などの物体の位置を特定する物体検出[1,2,3,4,5]や画素ごとに物体の種類を推定するセマンティックセグメンテーション[6,7,8,9]などが広く用いられている．特に物体検出は，信号機認識[10]や歩行者の移動を予測する経路予測[11,12,13]などの様々な応用方法が存在するため，重要な技術となっている．

近年の深層学習技術，とりわけ畳み込みニューラルネットワーク (Convolutional Neural Network; CNN) の発達により，CNN ベースの物体検出手法が多数提案されており，高い検出性能を実現している．これらのネットワーク構造は，あらかじめ人が手動で決定，構築している．自動運転における物体検出では，車両の近辺だけでなく遠方の物体を検出することが重要となるが，これらの手法は主に一般物体の検出[25,26]によって評価をしている．そのため，これらのネットワーク構造が自動運転を想定した車載カメラ画像や画像中の遠方物体を検出するために有効な構造であるかという問題については，十分な議論がなされていない．

そこで，本研究では，遠方物体を対象とした物体検出を目的とする．本稿では，与えられた画像認識問題に対する最適なネットワークの構造を自動的に探索する Neural Architecture Search (NAS)[14]およびパラメータ探索に基づき，画像サイズや学習率等のネットワークや学習に関するパラメータを探索する．手動ではなく適切なパラメータを自動的に探索し選択することにより，遠方物体検出に最適なネットワーク構造を学習する．評価実験では，車載カメラ画像を用いたデータセットを用いて評価を行い，提案手法が遠方物体に対して高い精度で検出できることを示す．

1) 中部大学(487-8501 愛知県春日井市松本町 1200)

2. 関 連 研 究

2.1. 物体検出

物体検出は画像認識分野において広く扱われているタスクの一つである．従来の物体検出では，人が手動で設計した特徴量と識別機を組み合わせた検出手法が提案されている．Dalal ら[15]は，歩行者検出のために Histograms of Oriented Gradients (HOG) という特徴量を提案しており，HOG 特徴量と Support Vector Machine (SVM) を組み合わせることにより歩行者検出を実現している．

近年の CNN 技術の発達により，CNN を用いた物体検出手法が多数提案されている[1,2,3,4,5]．これら CNN の物体検出手法では，特徴量を人手で設計することなく，ネットワークの学習を通じて問題に対する最適な特徴を抽出するパラメータをパラメータ群を獲得することが可能である．これにより，従来の手法に比べ高い検出精度を実現している．

CNN の物体検出手法には，two-stage および single-stage の 2 種類のアプローチが存在している．Two-stage の物体検出手法では，あらかじめ入力された画像に対して物体の候補となる領域を決定するネットワークと各候補領域がぞくするクラスを推定するネットワークの二段階の処理によって検出が

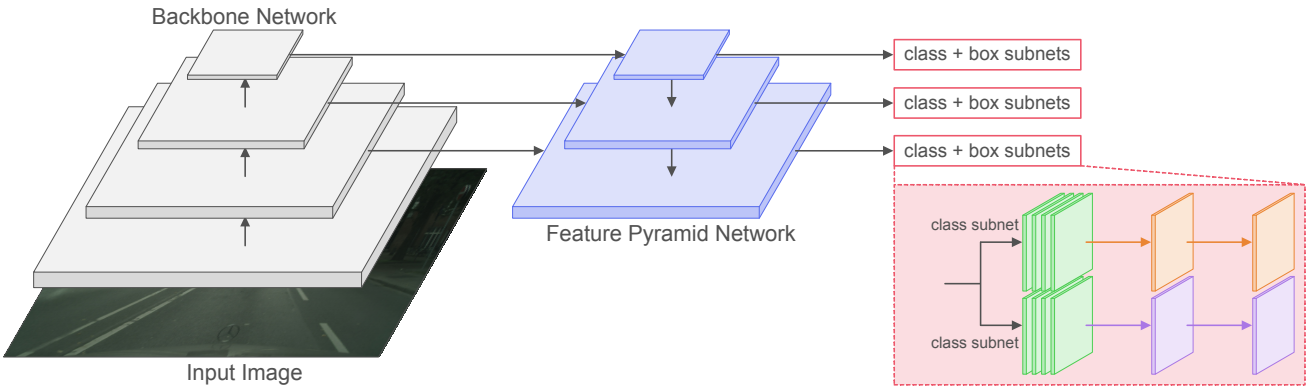

Fig. 1 The overview of the proposed network architecture

実現される．代表的な手法として，Faster R-CNN [5] が存在する．Faster R-CNN では， Region Proposal Network (RPN) により物体の候補領域を抽出し，抽出した候補領域に対応する特徴マップの情報から物体クラスを推定することで物体検出を実現する．

一方，Single-stage の物体検出手法は，単一のネットワークにより特徴抽出から検出までを一貫して行うアプローチである．Single-stage の代表的な物体検出手法としては，Single Shot MultiBox Detector (SSD) [4] や You Only Look Once (YOLO) [1,2,3] などが存在する．Single-stage は two-stage アプローチに比べ，ネットワーク構造がシンプルなことから処理時間が短いという特徴があり，自動運転などのリアルタイム性が求められるアプリケーションに有用と考えられる．その一方で，two-stage と比べ検出精度が低いという問題が存在する．その問題を改善するために，複数の解像度の特徴マップを用いて検出を行う Feature Pyramid Network (FPN) [16] や，学習時に使用する誤差関数の改善 [17] などの様々な工夫が提案されている．

これらの CNN に基づく検出手法では，ネットワーク構造を人が手動で設計しており，自動運転時の遠方物体の検出に適切であるかという問題については，十分な議論がなされていない．本研究では，遠方物体検出に適切なネットワーク構造を探索し，その結果について議論する．

2.2. Neural Architecture Search

人が手動でネットワーク構造を設計することなく，与えられたデータや問題に対して適切なネットワーク構造を探索する Neural Architecture Search (NAS) が盛んに研究されている [14]．ネットワーク構造の探索方法には，進化的計算アルゴリズム [18] や強化学習 [19] を用いた手法が提案されている．

MnasNet [21] は，モバイルデバイスのような計算性能の低い条件下で一般物体認識の画像認識タスクを動作させることを想定し，軽量かつ高精度な認識を実現するために提案されたネットワーク探索手法である．この手法では，認識性能および処理時間 (latancy) から報酬関数を定義し，強化学習により探索を行う．これにより，精度だけでなく，使用するデバイスの処理性能に応じた適切なネットワーク構造を決定することを可能としている．

物体検出のための NAS の代表的な手法として NAS-FPN [20] が存在する．NAS-FPN では，畳み込み処理により得られた複数の特徴マップを段階的に統合する FPN の処理を一般化し，複数の特徴マップの最適な統合方法を自動的に探索している．この手法では，一般物体のデータセット [26] に対して有効性を示しているが，自動運転のための車載カメラ画像に対する物体検出や遠方物体の検出に対する最適なアーキテクチャは十分に探索されていない．本研究では，ネットワーク構造を探索することで，自動運転時における遠方物体を検出するための有用なネットワーク構造を探索する．

3. 手 法

本節では，車載カメラ画像における遠方物体検出のための最適なネットワーク探索手法について述べる．

3.1. ネットワーク構造

本研究で用いるネットワーク構造を図 1 に示す．ここでは，物体検出のネットワーク構造として single-stage アプローチの FPN [16] を使用する．FPN では，バックボーンと呼ばれる CNN により入力画像から特徴抽出を行う．その後，バックボーンから抽出した複数の異なるサイズの特徴マップを FPN へ入力する．FPN では，バックボーンより得られた特徴マップを小さいものから段階的に統合し，複数の特徴マップを生成する．これにより，様々な物体スケールに対応した検出を構成度に実現することが可能となる．FPN より得られた特徴マップを class subnet および box subnet へと入力し，物体クラスの推定結果およびバウンディングボックス回帰の推定結果を得ることにより物体検出を行う．

FPN のうち，入力画像から特徴抽出を行うバックボーンネットワークとして，MnasNet-A1 [21] を用いる．MnasNet-A1 は，前述のようにモバイルデバイス等での CNN の使用を目的に探索された結果獲得されたネットワークであり，認識精度の維

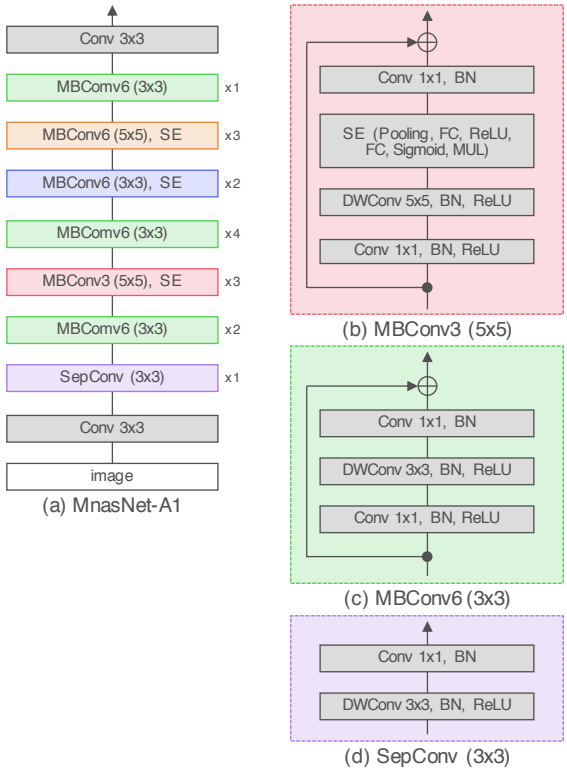

Fig. 2 The detailed network architecture of MnasNet-A1. "DWConv" indicates a depth-wise convolution, and "BN" indicates a batch normalization, "SE" indicates a squeeze-and-excitation module.

持をしつつ処理時間の高速化を実現したネットワークである．図 2 に MnasNet-A1 のネットワーク構造を示す．図 2 より，従来の CNN とは異なり，複数の異なるカーネルサイズの畳み込み層が不規則に存在していることがわかる．

3.2. パラメータ探索

前述のネットワークにおいて，最適なパラメータを探索することで，高精度な物体検出を実現する．提案手法では，学習率，weight decay，入力画像サイズ，FPN のチャンネル数，学習回数の最適化を行う．

パラメータの組み合わせは膨大であり，最適解を解析的に求めることは難しい．本研究では，optuna [22] を用いて探索を行う．探索したパラメータのうち，最も精度が高いパラメータを採用する．

4. 実 験

本節では，提案したネットワーク構造の探索手法の有効性を評価する．

4.1. データセット

実験には，Cityscapes dataset [23] および Berkley Deep Drive 100K (BDD100K) dataset [24] を用いる．Cityscapes dataset はドイツ国内の都市の市街地を走行し撮影したデータセットである．また，BDD100K はアメリカ国内の道路を走行し撮影したデータセットである．Cityscapes dataset の画像サイズは 2048x1024 [pixels]，BDD100K dataset の画像サイズは 1280x740 [pixels] である．Cityscapes Dataset における学習サンプル数は 2,975，評価サンプル数は 500 である．BDD100K dataset における学習サンプル数は 70,000，評価サンプル数は 10,000 である．また，検出を行う物体クラスは person，car，bike，traffic light の 4 種類である．

Table 1 Search range of hyperparameters

Parameter	Min. value	Max value
Learning rate	0.002	0.005
Weight decay	10^{-7}	10^{-4}
Image size	500×833	800×1333
# of channels of FPN	128	256
# of training steps	160,000	520,000

4.2. 評価

本実験では，Average Precision (AP) を用いて物体検出の精度を評価する．このとき，各データセットの物体サイズを small (25 [pixels] 未満)，medium (25 [pixels] 以上 35 [pixels] 未満)，large (35 [pixels] 以上) に分類し，物体のサイズに応じた検出率の評価を行う．

4.3. 学習方法

ネットワークの学習時には，Cityscapes dataset および BDD100K dataset の学習サンプルを一度に用いてネットワークを学習する．このとき，パラメータ探索により決定された画像サイズにリサイズし使用する．

パラメータの探索は optuna で行う．探索するパラメータおよびその探索範囲を表 1 に示す．MnasNet-A1 の学習では，ミニバッチサイズも探索の対象としているが，本実験では学習時のミニバッチサイズを 2 に設定した．表 1 のパラメータを決定し，学習を行った際の BDD100K dataset の評価サンプルに対する性能が最も高くなるパラメータの組み合わせを探索する．パラメータ探索およびネットワークの学習環境として，Quadro P5000 (16GB) を 60 枚使用し，約 200 回の探索を行なった．

4.4. 結果

4.4.1. mAP による定量評価

表 2 に，各データセットにおける mAP を示す． 4.1. 節に示したように，BDD100K の学習サンプル数が Cityscapes と比べて非常に多く，BDD100K における精度をパラメータ探索の基準としているためだと考えられる．そのため，Cityscapes と BDD100K を比較すると，BDD100K の方が mAP の値が高い．

また，物体サイズごとの精度を比較すると，どちらのデータセットにおいても，物体のサイズが小さい場合に精度が低

Table 2 Average precisions on Cityscapes and BDD100K datasets

Dataset	Average Precision				
	person	car	bike	traffic light	all (mAP)
Cityscapes (small)	0.087	0.211	0.027	0.289	0.153
Cityscapes (medium)	0.186	0.346	0.107	0.300	0.235
Cityscapes (large)	0.759	0.862	0.612	0.662	0.724
Cityscapes (all)	0.647	0.805	0.535	0.526	0.628
BDD100K (small)	0.276	0.510	0.127	0.614	0.382
BDD100K (medium)	0.356	0.463	0.157	0.468	0.361
BDD100K (large)	0.780	0.909	0.641	0.642	0.743
BDD100K (all)	0.693	0.841	0.552	0.730	0.704

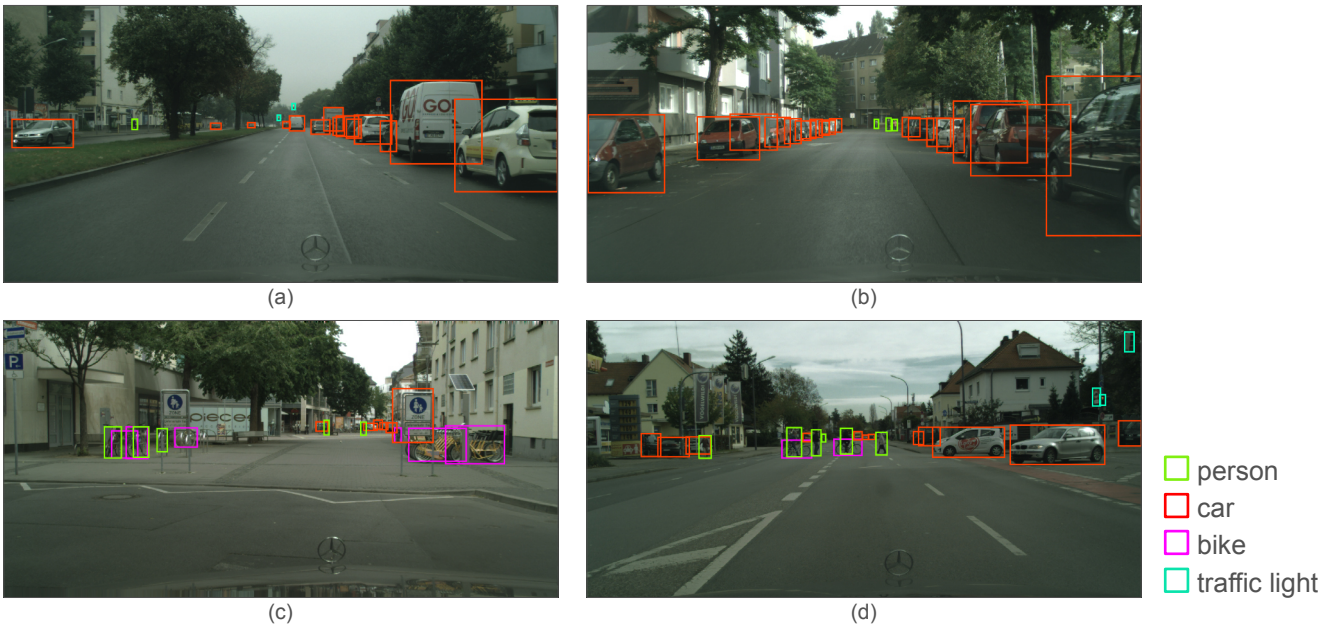

Fig. 3 Detection results on Cityscapes dataset

いことがわかる．特に，Cityscapes dataset においては，物体サイズが 35［pixels］以下の small，medium において，検出精度が著しく低下している．Citycapes dataset は画像サイズが 2048x1024［pixels］であり，BDD100K dataset と比較すると，相対的に物体サイズが小さい．そのため，BDD100K dataset よりも検出の困難な物体に対する評価を行なっていることに相当する．画素単位での物体サイズごとの評価だけでなく，実際の距離や物体サイズに応じた評価を行うことも今後の課題の一つである．

4.4.2. 検出結果例

図 3 に Cityscapes dataset における検出結果例を示す．図 3(a，b) では遠方の歩行者や信号機などの遠方の小さな物体も適切に検出している．一方，図 3(c) は画像左のポールを person や bike と検出しており，図 3(d) の右では標識を traffic light と検出している．このように，物体のスケールにかかわらず，誤検出が多く発生しているために，表 2 に示す AP に影響していると考えられる．

図 4 に BDD100K dataset における検出結果例を示す．BDD100K dataset では，遠方の物体においても適切に検出できていることがわかる．図 4(b) は，高速道路を走行中のシーンである．このような状況では，走行速度が速いため，市街地よりも遠方の物体を検出することが重要となる．このようなシーンにおいても，適切に検出できていることがわかる．また，図 4(c，d) は夜間の走行シーンであり，より検出が困難な状況となっている．探索により得られたネットワークでは，夜間のシーンであっても，遠方の小さな物体を適切に検出できている．これより，獲得したネットワークにより遠方物体を適切に検出していることがわかる．

4.4.3. 探索により獲得したパラメータ

表 3 にパラメータ探索により獲得したパラメータを示す．画像サイズおよび，FPN の特徴マップのチャンネル数は探索範囲のうち，最大の値が選択されている．これより，遠方の物体を検出するためには，大きなサイズの画像を入力として用いる必要がある．しかし，本実験での探索回数は 200 回であり，探索するべきパラメータ空間に対して探索回数が不十分だと考えられる．また，今回は 800x1333 を最大の画像サイズ

Fig. 4 Detection results on BDD100K dataset

Table 3 Obtained hyperparameters

Parameter	Value
Learning rate	0.003
Weight decay	1.02^{-6}
Image size	800×1333
# of channels of FPN	256
# of training steps	440,000

と設定したが，より大きな画像サイズでの探索も行う必要がある．そのため，探索回数を増やすことで，より良いパラメータを獲得する必要がある．

5. ま と め

本稿では，自動運転における車載カメラ映像に対する遠方物体検出のための最適なネットワーク構造を探索するための手法を提案した．提案手法は，FPN と NAS に基づき，自動的に探索を行う．Cityscapes Dataset および BDD100K Dataset を用いた評価実験により，遠方の物体を適切に検出することを可能とした．

今回の実験では，30 枚の GPU を用いて探索を行ったが，より最適な値を求めるためには探索回数を増やす必要がある．また，バックボーンネットワークおよび FPN のアーキテクチャ探索も同時に行うことで，より高精度な検出を実現することが可能と考えられる．しかし，現在の探索アルゴリズムでは膨大な時間を要することから，これらすべての探索を十分な回数行うことは困難である．そのため，今後の予定として，短時間でより最適な構造を模索するために，探索手法を改善することなどが挙げられる．

謝 辞

本研究は，総合科学技術・イノベーション会議の戦略的イノベーション創造プログラム（SIP）第 2 期/自動運転（システムとサービスの拡張）「自動運転技術（レベル 3，4）に必要な認識技術等に関する研究」（管理法人：NEDO）によって実施されました．

参 考 文 献

(1) Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi : You Only Look Once: Unified, Real-Time Object Detection, in Computer Vision and Pattern Recognition (2016)

(2) Joseph Redmon, Ali Farhadi : YOLO9000: Better, Faster, Stronger, in Computer Vision and Pattern Recognition (2017)

(3) Joseph Redmon, Ali Farhadi : YOLOv3: An Incremental Improvement, arXiv preprint, arXiv:1804.02767 (2018)

(4) Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, Alexander C. Berg : SSD: Single Shot MultiBox Detector, in European Conference on Computer Vision (2016)

(5) Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun : Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, in Neural Information Processing Systems (2015)

(6) Vijay Badrinarayanan, Alex Kendall, Roberto Cipolla : SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 39, No. 12, pp. 2481-2495 (2017)

(7) Evan Shelhamer, Jonathan Long, Trevor Darrell : Fully Convolutional Networks for Semantic Segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 39, No. 4, pp. 640-651 (2017)

(8) Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, Jiaya Jia : Pyramid Scene Parsing Network, in Computer Vision and Pattern Recognition (2017)

(9) Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, Hartwig Adam : Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation, in European Conference on Computer Vision (2018)

(10) Xi Li, Huimin Ma, Xiang Wang, Xiaoqin Zhang : Traffic Light Recognition for Complex Scene With Fusion Detections, IEEE Transactions on Intelligent Transportation Systems, Vol. 19, No. 1, pp. 199-208 (2018)

(11) Julian Francisco, Pieter Kooij, Nicolas Schneider, Fabian Flohr, Dariu M. Gavrila : Context-Based Pedestrian Path Prediction, in European Conference on Computer Vision (2014)

(12) Julian F. P. Kooij, Fabian Flohr, Ewoud A. I. Pool, Dariu M. Gavrila : Context-Based Path Prediction for Targets with Switching Dynamics, International Journal of Computer Vision, Vol. 127, pp. 239-262 (2019)

(13) Apratim Bhattacharyya, Mario Fritz, Bernt Schiele : Long-Term On-Board Prediction of People in Traffic Scenes under Uncertainty, in Computer Vision and Pattern Recognition (2017)

(14) Thomas Elsken, Jan Hendrik Metzen, Frank Hutter : Neural Architecture Search: A Survey, Journal of Machine Learning Research, Vol. 20, No. 55, pp. 1-21 (2019)

(15) Navneet Dalal, Bill Triggs : Histograms of oriented gradients for human detection, in Computer Vision and Pattern Recognition (2005)

(16) Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, Serge Belongie : Feature Pyramid Networks for Object Detection, in Computer Vision and Pattern Recognition (2017)

(17) Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, Piotr Dollár : Focal Loss for Dense Object Detection, in International Conference on Computer Vision (2017)

(18) Esteban Real, Alok Aggarwal, Yanping Huang, Quoc V Le : Regularized Evolution for Image Classifier Architecture Search, in Association for the Advancement of Artificial Intelligence (2019)

(19) Barret Zoph, Quoc V. Le : Neural Archtecture Search With Reinforment Learning, in International Conference on Learning Representations (2017)

(20) Golnaz Ghiasi, Tsung-Yi Lin, Ruoming Pang, Quoc V. Le : NAS-FPN: Learning Scalable Feature Pyramid Architecture for Object Detection, in Computer Vision and Pattern Recognition (2019)

(21) Mingxing Tan, Bo Chen, Ruoming Pang, Vijay Vasudevan, Mark Sandler, Andrew Howard, Quoc V. Le : Mnasnet: Platform-aware neural architecture search for mobile, in Computer Vision and Pattern Recognition (2019)

(22) Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, Masanori Koyama : Optuna: A Next-generation Hyperparameter Optimization Framework, arXiv preprint, arXiv:1907.10902 (2019)

(23) Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, Bernt Schiele : The Cityscapes Dataset for Semantic Urban Scene Understanding, in Computer Vision and Pattern Recognition (2016)

(24) Fisher Yu, Wenqi Xian, Yingying Chen, Fangchen Liu, Mike Liao, Vashisht Madhavan, Trevor Darrell : BDD100K: A Diverse Driving Video Database with Scalable Annotation Tooling, arXiv preprint, arXiv:1805.04687 (2018)

(25) Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, Andrew Zisserman : The PASCAL Visual Object Classes (VOC) Challenge, International Journal of Computer Vision, Vol. 88, pp. 303-338 (2010)

(26) Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, C. Lawrence Zitnick : Microsoft COCO : Common Objects in Context, in European Conference on Computer Vision (2014)