一般社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

HCGシンポジウム2019
HUMAN COMMUNICATION GROUP SYMPOSIUM 2019
HCG2019-C-4-4

複数手法の統合によるセマンティックセグメンテーションのクラス拡張

汪　邱晟† 川西　康友†† 出口　大輔††† 井手　一郎†† 村瀬　洋††

† 名古屋大学大学院情報科学研究科 〒 464–8601 愛知県名古屋市千種区不老町
†† 名古屋大学大学院情報学研究科 〒 464–8601 愛知県名古屋市千種区不老町
††† 名古屋大学情報戦略室 〒 464–8601 愛知県名古屋市千種区不老町
E-mail: †wangq@murase.m.is.nagoya-u.ac.jp, ††{kawanishi,ide,murase}@i.nagoya-u.ac.jp,
†††ddeguchi@nagoya-u.jp

あらまし 近年，自動運転技術や拡張現実（AR）の発展に伴い，周囲環境を詳細に理解するための技術が求められている．その中でもセマンティックセグメンテーションは画像の全画素に対してクラス分類が可能な技術であることから，様々な応用において有力な技術として注目を集めている．深層学習を利用したセマンティックセグメンテーション手法の学習には大量の学習データが必要であるものの，自力で新たなデータセットを用意することがとても難しいため，既存のデータセットを活用しつつ，更に多くのクラスを認識できることが望ましい．本発表では，既存のデータセットを活用することで，認識できるクラスを任意に拡張できるセマンティックセグメンテーション手法を提案する．評価実験の結果，クラスの拡張に成功し，提案手法の有効性を確認した．
キーワード セマンティックセグメンテーション, クラス拡張

Class Augmentation For Semantic Segmentation by Integrating Multiple Methods

Qiusheng WANG†, Yasutomo KAWANISHI††, Daisuke DEGUCHI†††, Ichiro IDE††, and Hiroshi MURASE††

† Graduate School of Information Science, Nagoya University
Furo–cho, Chikusa–ku, Nagoya–shi, Aichi, 464-8601 Japan
†† Graduate School of Informatics, Nagoya University
Furo–cho, Chikusa–ku, Nagoya–shi, Aichi, 464-8601 Japan
††† Information Strategy Office, Nagoya University
Furo–cho, Chikusa–ku, Nagoya–shi, Aichi, 464-8601 Japan
E-mail: †wangq@murase.m.is.nagoya-u.ac.jp, ††{kawanishi,ide,murase}@i.nagoya-u.ac.jp,
†††ddeguchi@nagoya-u.jp

Abstract Along with the development of autonomous driving and Augmented Reality (AR), we need technologies that can help understanding surrounding environment better. Semantic segmentation, which applies classfication to every single pixel in an image, is gathering attention as a powerful tool in many fields. While semantic segmentation with deep learning techniques requires a huge amount of data for training models, building a dataset on our own is costly. Therefore, we want to use existing datasets instead, while also being able to recognize more classes than a single dataset provides. In this report, we propose a semantic segmentation method that can augment the classes on demand using existing datasets. We confirmed the effectiveness of the proposed method with the success of class augmentation in an evaluation experiment.
Key words Semantic segmentation, class augmentation

1. はじめに

セマンティックセグメンテーションとは，画像を画素ごとに属するクラスのラベルを割り当てる問題である．これは従来の画像認識タスクと比べて，映っている物体のクラスの認識に加え，形状を画素単位で正しく把握する問題設定となっている．そのため問題そのものの難易度は高く，また計算量も多い．近年では，画像認識における基本的かつ重要な課題になっている．また，自動運転に不可欠な周囲環境の把握や，スマートフォン等のカメラを活用した拡張現実（AR）技術などの様々な応用においても注目を浴び始めている．

セマンティックセグメンテーションは，深層学習に基づく手法が多いが，出力次元数が大きいことから学習に大規模なデータセットが必要不可欠である．しかしセマンティックセグメンテーションのデータセットの構築には撮影とアノテーション作業が必要であり，これらは人的コストの面から簡単に用意できない．公開されている一般的なセマンティックセグメンテーションデータセットとしては，Cityscapes [13]，Microsoft COCO [14]，ADE20K [15] [16]，ApolloScape [17] などが挙げられ，収録するクラス数が数十から百を越えるものまで様々である．また，数千から数万枚の画像が含まれる大規模なデータセットが多い．

実際の応用を考えた場合，セマンティックセグメンテーションのクラス設定は課題に応じて適切に定める必要があるが，公開されている既存のデータセットではクラスの種類やアノテーションの基準が各々異なり，1 つのデータセットのみでは必要なクラスが含まれていない可能性が高い．そのため，特定の既存データセットが必ずしも需要と合致するわけではない．しかし独自のデータセットは簡単に用意できないことから，既存のデータセットを組み合わせて目的を達成する技術が強く求められている．

そこで本発表では，必要に応じて認識するクラスを増やせるセマンティックセグメンテーションの実現を目的とし，複数の既存データセットを併用することでクラスの種類を拡張可能な汎用的な手法を提案する．一般に公開されているデータセットのクラス定義が各々異なるだけでなく，意味が重複した類似クラスも存在する．そのため，単純な組み合わせでは深層学習に利用困難である．提案手法では，各データセットで学習済みのモデルによる出力の共起性に基づく共起ルール，類似クラスに対してどちらのデータセットのアノテーションを優先するかの制約ルールの 2 つのルールからなる統合法を提案することで，これらの問題点に対処する．この統合法により，複数のデータセットによるクラス種類の拡張を可能にする．

2. 関連研究

関連研究として，既存のセマンティックセグメンテーション手法を紹介する．一般に，セマンティックセグメンテーションの方法として，Instance-level と Pixel-level の 2 種類がある．両者の違いを図 2 に示す．

Instance-level 手法は，まず従来の物体認識のように画像

図 1: セマンティックセグメンテーションの例（ADE20K [15]）

に映っている対象物の種類や位置を認識したうえで，各物体の領域に対してセグメンテーションを行う枠組みである．従来の物体認識から発展させたような方法であるため，物体認識手法から発展させたもの [1] [2] [3] が多い．その中でも，Mask R-CNN [1] は Fast/Faster RCNN [18] [19] にセグメンテーション層を追加して，さらに Feature Pyramid Network (FPN) [20] を使用することで，異なるスケールの情報を捕捉できるようにした簡単かつ有効な手法として有名である．PANet [2] は Mask R-CNN [1] に対してネットワーク内の情報伝達経路を改良し，さらに adaptive feature pooling を導入することでセグメンテーションの精度向上に成功している．特徴としては物体を認識してからセグメンテーションを行うため，人間の画像認識に近く，結果を利用しやすい，また物体が重なっている時に各々を区別することができるが，検出できていない物体や認識できる物体以外の領域に対してはセグメンテーションができないといった課題がある．

Pixel-level 手法は，物体 1 つ 1 つを認識することなく，入力画像をそのまま全画素に対してセグメンテーションを行う枠組みである．そのため，Instance-level 手法と比べてより密な構造情報が得られるが，重なっている物体を区別することができない．また，Instance-level 手法と比べて計算量が多くなる傾向がある．Pixel-level 手法では，特に複数スケールでのコンテキストを捕捉可能なモジュールと骨格となるネットワークの構造設計が重要である．画素単位のクラス分類タスクにおいて，コンテキスト情報が重要であることは複数の研究 [22] [23] [24] [25] によって指摘されている．PSPNet [9] は spatial pyramid pooling を異なるグリッドサイズで行うことでコンテキスト情報を取得している．DeepLab [5] [6] では代わりに異なる間隔での atrous convolution [26] [27] [28] [29] を使用している．骨格のネットワークは設計の改良に伴い，従来の AlexNet [30]，VGG [31]，ResNet [32] などから最近の ResNext [33]，DenseNet [34]，Xception [35] まで，精度や速度面で大きく進歩している．また最近では，ネットワークアーキテクチャの構築にメタ学習を適用する試みもある [36] [37]．DPC [10] はネットワークの一部にメタ学習によって得られた最適な構造を使用しており，また Auto-Deeplab [11]

入力画像　　Pixel-level　　Instance-level

図 2: Pixel-level 手法と Instance-level 手法の違い．Pixel-level 手法では同じ種類の物体に対して同じラベルを割り当てているのに対して，Instance-level 手法では違う物体と見なして，違うラベルを割り当てている．

はネットワーク全体をメタ学習することで，人間の設計に縛られない最適なネットワーク構造の探索を試みている．

また，最近では Panoptic Segmentation [12] という，Pixel-level のセグメンテーションに加え，各物体を区別できるようにラベル付けする Instance-level と Pixel-level を組み合わせた方法が提案されている．UPSNet [3] は Mask R-CNN [1] に deformable convolution [21] を使用したセグメンテーション層を追加し，Pixel-level のセグメンテーションにも対応できるようにしたうえで，Panoptic Head によって Pixel-level と Instance-level の出力を統合することで Panoptic Segmentation を可能にしている．

以上，これらの手法はすべて事前に与えられたデータセットのクラス定義に従って学習を行うフレームワークである．そのため，そのままでは任意のクラスに拡張させることはできない．

3. 提 案 手 法

本節では，提案手法の詳細について述べる．提案手法では，事前学習で生成したルールに基づき，各データセットで学習済みモデルの出力を統合することで，クラス拡張を行う．

セマンティックセグメンテーションには Pixel-level 手法の学習済みモデルを利用する．事前準備として，基のデータセット（以下，主データセットと呼ぶ）と，拡張したいクラスが含まれているデータセット (以下，副データセットと呼ぶ) を用意する．主データセットのクラスに，副データセットから必要なクラスを取り出して追加したものが，出力結果のクラスの集合となる．また，主・副データセット各々で学習させたモデル（以下，それぞれ主モデル・副モデルと呼ぶ）を用意する．この時，モデルの精度差による結果の偏りを抑制するため，なるべく同じ手法を 2 つのデータセットで学習させたモデルを使用するのが望ましい．

処理の手順は以下の通りである．

（1） 事前学習として副データセットの画像を主モデル・副モデルに入力し，それぞれの出力と真値との共起性に着目した共起ルールを得る．このとき，データセット間の類似クラスを適切に扱うためのルール (以下, 制約ルールと呼ぶ) も合わせ，統合ルールとする．

（2） 入力画像を主モデル・副モデルに入力し，それぞれの出力を得る．

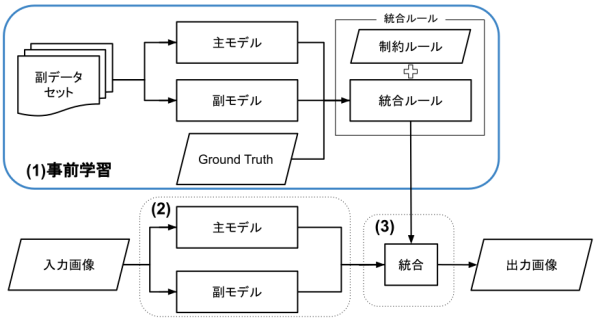

図 3: 提案手法の処理手順

（3） 統合ルールに基づきそれぞれの出力を統合することで，最終の出力結果を得る．提案手法の処理手順を図 3 に，統合の概念図を図 4 に示す．

以下，処理の詳細についてそれぞれ説明する．

3.1 共起ルール

主データセットのクラスを $C_i \in S_{\mathrm{main}}(i = 1, 2, ..., m)$，副データセットのクラスを $C_j \in S_{\mathrm{sub}}(j = 1, 2, ..., n)$ と表す．ある画像の 1 画素に対して，主モデルの出力を C_i，副モデルの出力を C_j，そして真値のクラスを $C_{\mathrm{gt}}(C_{\mathrm{gt}} \in S_{\mathrm{sub}})$ とした場合，共起ルールは式 (1) のように表せる．

$$f(C_i, C_j) = C_{\mathrm{gt}} \tag{1}$$

式 (1) は，1 枚の画像の各画素に対して，2 つのモデルから出力されたラベルと真値のラベルの組み合わせを記録することで得られる．具体的には，主モデルと副モデルを副データセットの Validation 画像すべてに適用し，ラベルの組み合わせの数（該当する画素の数）を集計する．集計結果から，主モデル・副モデルの出力のある組み合わせ (C_i, C_j) の中で最も出現頻度が高い真値のラベル C_{gt} を選択する．これを式 (1) のようなルールとする．

実際の統合時においては，ある画素に対して主モデルが C_i，副モデルが C_j を出力した場合，式 (1) のルールに従い C_{gt} を出力する．

ただし，上述のようなルールでは，各データセットのラベル付け基準の違いをうまく吸収することはできない．例えば，

- 自転車に乗っている人を，主データセットではそれぞれ人間と自転車のクラスとして扱っているのに対し，副データセットはそれらを合わせてライダーとして扱っている場合，出力結果もライダーとなる．
- 同じ意味を持つクラス同士の場合，主データセットのクラスが，副データセットのクラスに書き換えられてしまうといった，単なるクラス名の変更が起きる．

これらの問題を解決し，また本研究の目標である，必要に応じてクラス拡張を実現するためには，追加の制約ルールを設けることが必要である．

3.2 制約ルール

セマンティックセグメンテーションのデータセットでは，同じまたは似たような意味をもつクラスが存在する．例として，

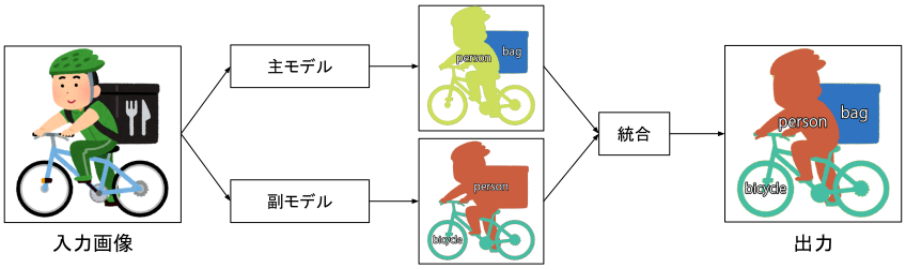

図 4: 統合処理の概念図．主データセットではライダーという単一クラスとして扱っていた部分を，副データセットからクラスを追加することで人，バッグ，自転車の 3 クラスに区別できるようになる．

表 1: Cityscapes [13] と ADE20K [15] の類似クラス（一部）

Cityscapes	ADE20K
road	road, route
building	building, edifice, house, skyscraper
pole	pole, column, pillar

表 2: 検証実験結果

手法	Cityscapes mIOU(%)	拡張クラス **mIOU**(%)	mIOU(%)
DPC (Cityscapes)	64.2	**0.0**	57.1
提案手法	57.8	**19.6**	53.6

Cityscapes [13] と ADE20K [15] を比較した時の類似クラスを表 1 に示す．

このような類似するクラスに対しては，主データセットのクラスを優先するように制約ルールを設ける．また，追加するクラスを選択するルールも必要に応じて追加する．

主データセットの 1 クラス C_i に対して，副データセットに含まれる類似クラスの集合を $S_{C_i} \subset S_{\mathrm{sub}} (i = 1, 2, ..., m)$ と表す．また，副データセットから追加するクラスの集合を $S_{\mathrm{add}} \subseteqq S_{\mathrm{sub}}$，拡張されたクラスの集合を $S_{\mathrm{aug}} = S_{\mathrm{main}} \cup S_{\mathrm{add}}$ とする．制約ルールは式 (2) のように表せる．

$$g(C_i, C_j) = \begin{cases} C_i & \text{if } f(C_i, C_j) \in S_{C_i}, \\ C_i & \text{if } f(C_i, C_j) \notin S_{\mathrm{aug}}, \\ f(C_i, C_j) & \text{otherwise.} \end{cases} \quad (2)$$

4. 実　　験

提案手法の有効性を検証するため，実験を行う．

4.1 データセットとモデル

本実験では，主に車載カメラでの用途を想定し，主データセットに Cityscapes [13]（19 クラス）を用意した．また，副データセットに ADE20K [15]（150 クラス）を用意した．クラス拡張では，ADE20K [15] の 150 クラスから Cityscapes [13] と重複するクラスを取り除いた全クラスを対象とした．事前学習の画像として，ADE20K の Validation データセットに含まれる 2,000 枚を使用した．

主モデルとしては DPC [10] を Cityscapes で学習させたものを使用し，副モデルとして DeepLabv3+ [7] を ADE20K で学習させたものを用いた．

また，評価用のデータに関しては，Cityscapes [13] の Validation 画像から 5 枚を選び，元の真値に基づいて，拡張されたクラス集合のアノテーションを追加した．

4.2 評 価 方 法

比較手法として，4.1 節で述べた主モデルをそのまま使用した（クラス拡張なし）．評価用のデータに比較手法と提案手法を適用し，全体の mean-IOU を計算して評価した．

4.3 実験結果および考察

実験結果を表 2 に，実際の出力結果を図 5 に示す．比較手法と比べて，提案手法では拡張クラスの mIOU が上昇していることから，クラス拡張に成功していることがわかる．ただし，比較手法と比べて，Cityscapes [13] のクラスでの mIOU が若干低下した．これは，主に物体の輪郭部分の精度が若干低下したためである．考えられる原因としては，本実験では副モデルを DeepLabv3+ [7] で代用しているため，DPC [10] と比べてセグメンテーション精度が不十分であったことが考えられる．また，提案手法では 1 画素ごとに処理を行っていることから，画像に含まれるコンテキスト情報を捉えることができないため，物体を物体として認識できないことが原因である．さらに，主・副モデルに同じ（実際は似たような）手法を使用しても，異なるデータセットで学習したモデルではセグメンテーションの精度差が存在するため，これも結果に影響していると考えられる．特に ADE20K [15] には合計 150 のクラスが含まれていることから，問題の難易度が高く，19 クラスしかない Cityscapes [13] の学習済みモデルと比べて精度が低い．実際の出力結果でも，同じ車両のはずが車輪部分だけバスとして認識されるようなこともしばしばである．このことから，副データセットにはクラス数が少ないデータセットを使用すれば，精度低下を抑えられると考えられる．

5. む　す　び

本研究では，セマンティックセグメンテーションに対して，既存のデータセットを利用して，認識できるクラス数を任意に拡張できるようにする手法を提案した．また実験により，手法の有効性を確認した．しかし物体の輪郭部分の精度が低いなどの

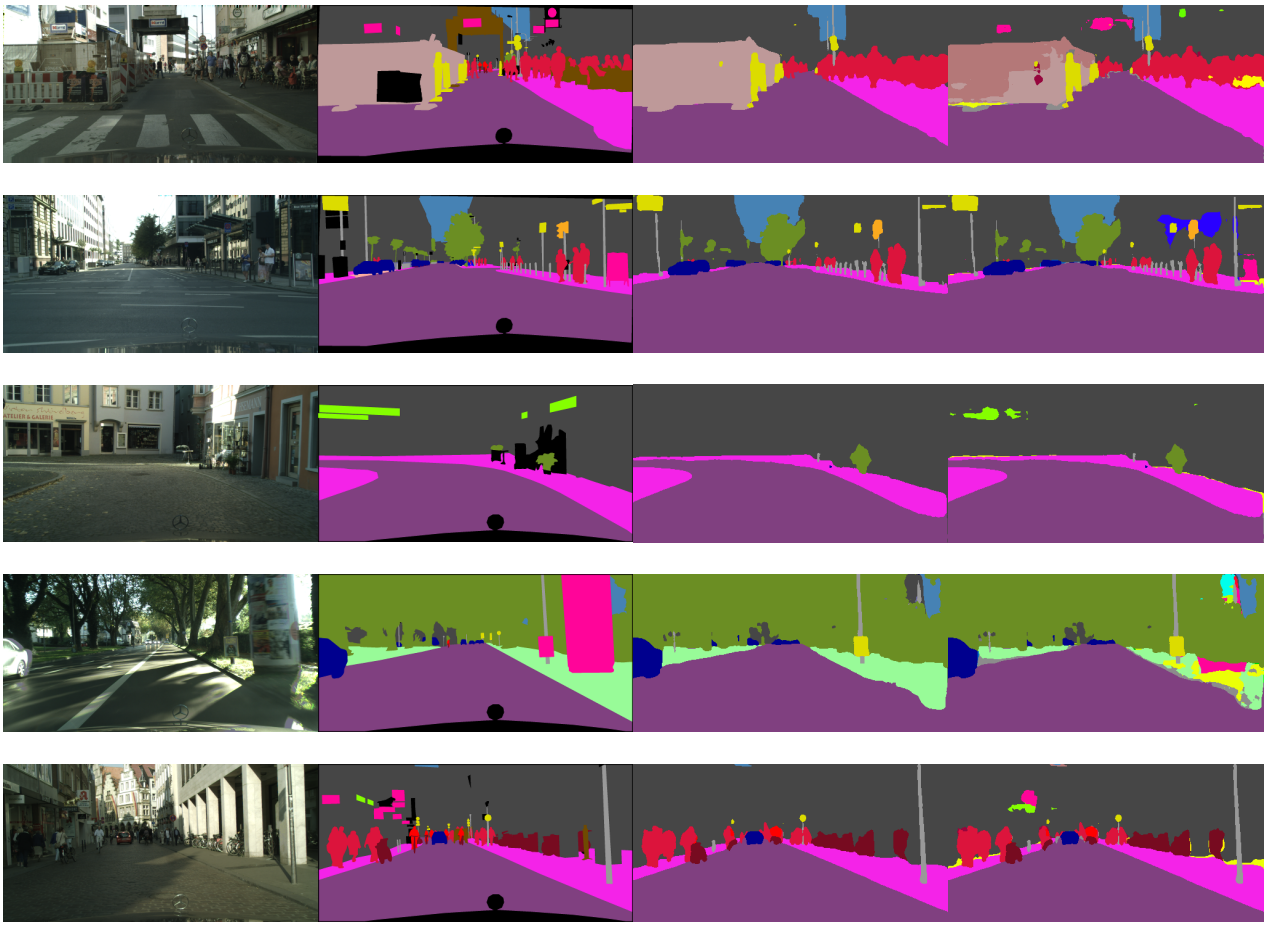

図 5: 出力結果例．看板などのクラスの拡張に成功していることがわかるが，副モデルの精度不足により，小さい看板などが検出できていない．

問題も残っている．今後は，輪郭部分の精度向上や，クラスラベルの出力の代わりにクラス確率を使うなど，手法の改良を検討していきたい．

謝辞 日頃より熱心に御討論頂く名古屋大学村瀬研究室諸氏に深く感謝する．本研究の一部は、JSPS 科研費（17H00745）によるものである．

文　　献

[1] He, Kaiming, et al. “Mask R-CNN.” Proceedings of the IEEE International Conference on Computer Vision. 2017.

[2] Liu, Shu, et al. “Path Aggregation Network for Instance Segmentation.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

[3] Xiong, Yuwen, et al. “UPSnet: A Unified Panoptic Segmentation Network.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.

[4] Ronneberger, Olaf, et al. “U-net: Convolutional Networks for Biomedical Image Segmentation.” Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, Cham, 2015.

[5] Chen, Liang-Chieh, et al. “Deeplab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs.” IEEE Transactions on Pattern Analysis and Machine Intelligence 40.4: 834–848. 2017.

[6] Chen, Liang-Chieh, et al. “Rethinking Atrous Convolution for Semantic Image Segmentation.” arXiv preprint arXiv:1706.05587. 2017.

[7] Chen, Liang-Chieh, et al. “Encoder-decoder with Atrous Separable Convolution for Semantic Image Segmentation.” Proceedings of the European Conference on Computer Vision. 2018.

[8] Yu, Fisher, et al. “Multi-scale Context Aggregation by Dilated Convolutions.” arXiv preprint arXiv:1511.07122. 2015.

[9] Zhao, Hengshuang, et al. “Pyramid Scene Parsing Network.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.

[10] Chen, Liang-Chieh, et al. “Searching for Efficient Multi-scale Architectures for Dense Image Prediction.” Advances in Neural Information Processing Systems. 2018.

[11] Liu, Chenxi, et al. “Auto-deeplab: Hierarchical Neural Architecture Search for Semantic Image Segmentation.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.

[12] Kirillov, Alexander, et al. “Panoptic Segmentation.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.

[13] Cordts, Marius, et al. “The Cityscapes Dataset for Semantic Urban Scene Understanding.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[14] Lin, Tsung-Yi, et al. “Microsoft COCO: Common Objects in Context.” European Conference on Computer Vision. Springer, 2014.

[15] Zhou, Bolei, et al. “Scene Parsing Through ADE20K Dataset.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.

[16] Zhou, Bolei, et al. “Semantic Understanding of Scenes Through the ADE20K Dataset.” International Journal of Computer Vision 127.3: 302–321. 2019.

[17] Huang, Xinyu, et al. “The Apolloscape Dataset for Autonomous Driving.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2018.

[18] Girshick, Ross. “Fast R-CNN.” Proceedings of the IEEE International Conference on Computer Vision. 2015.

[19] Ren, Shaoqing, et al. “Faster R-CNN: Towards Real-time Object Detection with Region Proposal Networks.” Advances in Neural Information Processing Systems. 2015.

[20] Lin, Tsung-Yi, et al. “Feature Pyramid Networks for Object Detection.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.

[21] Dai, Jifeng, et al. “Deformable Convolutional Networks.” Proceedings of the IEEE international Conference on Computer Vision. 2017.

[22] He, Xuming, et al. “Multiscale Conditional Random Fields for Image Labeling.” Proceedings of the 2004 IEEE Conference on Computer Vision and Pattern Recognition, 2004.

[23] Shotton, Jamie, et al. “Textonboost for Image Understanding: Multi-class Object Recognition and Segmentation by Jointly Modeling Texture, layout, and Context.” International Journal of Computer Vision 81.1: 2–23. 2009.

[24] Chen, Liang-Chieh, et al. “Attention to Scale: Scale-aware Semantic Image Segmentation.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[25] Farabet, Clement, et al. “Learning Hierarchical Features for Scene Labeling.” IEEE Transactions on Pattern Analysis and Machine Intelligence 35.8: 1915–1929. 2012.

[26] Holschneider, Matthias, et al. “A Real-Time Algorithm for Signal Analysis with the Help of the Wavelet Transform.” Wavelets: Time-Frequency Methods and Phase Space. 286–297. 1990.

[27] Giusti, Alessandro, et al. “Fast Image Scanning with Deep Max-pooling Convolutional Neural Networks.” Proceedings of the IEEE International Conference on Image Processing. 2013.

[28] Sermanet, Pierre, et al. “Overfeat: Integrated Recognition, Localization and Detection Using Convolutional Networks.” arXiv preprint arXiv:1312.6229. 2013.

[29] Papandreou, George, et al. “Modeling Local and Global Deformations in Deep Learning: Epitomic Convolution, Multiple Instance Learning, and Sliding Window Detection.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.

[30] Krizhevsky, Alex, et al. “Imagenet Classification With Deep Convolutional Neural Networks.” Advances in neural information processing systems. 2012.

[31] Simonyan, Karen, et al. “Very Deep Convolutional Networks for Large-scale Image Recognition.” arXiv preprint arXiv:1409.1556. 2014.

[32] He, Kaiming, et al. “Deep Residual Learning for Image Recognition.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[33] Xie, Saining, et al. “Aggregated Residual Transformations for Deep Neural Networks.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.

[34] Huang, Gao, et al. “Densely Connected Convolutional Networks.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.

[35] Chollet, Franois. “Xception: Deep Learning with Depthwise Separable Convolutions.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017.

[36] Zoph, Barret, et al. “Learning Transferable Architectures for Scalable Image Recognition.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

[37] Liu, Chenxi, et al. “Progressive Neural Architecture Search.” Proceedings of the European Conference on Computer Vision. 2018.