UNIVERSITY OF TECHNOLOGY SYDNEY
Faculty of Engineering and Information Technology

Context-aware Image Semantic Segmentation

by

Ye Huang

A Thesis Submitted
in Fulfillment of the
Requirements for the Degree

Doctor of Philosophy

Sydney, Australia

2022

CERTIFICATE OF ORIGINAL AUTHORSHIP

I, Ye Huang, declare that this thesis, is submitted in fulfilment of the requirements for the award of Doctor of Philosophy, in the School of Electrical and Data Engineering, Faculty of Engineering and Information Technology at the University of Technology Sydney.

This thesis is wholly my own work unless otherwise referenced or acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This document has not been submitted for qualifications at any other academic institution.

This research is supported by the Australian Government Research Training Program.

Signature: Production Note:
Signature removed prior to publication.

Date: May-06-2022

ABSTRACT

Context-aware Image Semantic Segmentation

by

Ye Huang

Semantic segmentation is a fundamental task for computer vision applications. However, the existing solutions have many issues when handling difficult cases. This thesis develops three novel approaches which have improved the generalization ability of the existing solutions at significantly reduced computation costs. Extensive experiments conducted on multiple benchmark datasets have demonstrated the superior performance of the proposed approaches.

Scale-invariant: The state-of-the-art semantic segmentation solutions usually leverage different receptive fields via multiple parallel branches to handle objects of different sizes. However, employing separate kernels for individual branches degrades the generalization of the network to objects with different scales, and the computational cost increases with the increase of the number of branches. In this thesis, a novel network structure, namely *Kernel-Sharing Atrous Convolution (KSAC)*, is proposed, where branches with different receptive fields share the same kernel, i.e., letting a single kernel "see" the input feature maps more than once with different receptive fields.

Seamless dual attention: Spatial and channel attentions, modelling the semantic inter-dependencies in spatial and channel dimensions respectively, have recently been widely used for semantic segmentation. However, computing spatial attention and channel attention separately sometimes causes errors, especially in those difficult cases. In this research, a Channelized Axial Attention (CAA) is developed to seamlessly *integrate* channel attention and spatial attention into a single operation with negligible computation overhead. Furthermore, a novel grouped vectorization approach is developed to allow the proposed model to run with very little memory consumption without slowing down the computation.

Class-aware regularization: Recent segmentation methods utilizing class-level information in addition to pixel features have achieved notable success in boosting the accuracy of existing network models. However, the extracted class-level information was simply concatenated to pixel features, without being explicitly exploited to learn better pixel representation. Moreover, these approaches learn soft class centers based on coarse mask prediction, which is prone to error accumulation. Motivated by the fact that humans can recognize an object by itself no matter which other objects it appears with and aiming to use class-level information more effectively, a universal Class-Aware Regularization (CAR) approach is proposed to optimize the intra-class variance and inter-class distance during feature learning. Furthermore, the class center in the proposed approach is directly generated from ground truth instead of from the error-prone coarse prediction. The proposed CAR can be easily applied to most existing segmentation models and can largely improve their accuracy at no additional inference overhead.

Dissertation directed by Prof. Xiangjian He and Dr Wenjing Jia
School of Electrical and Data Engineering

Dedication

Dedicated to my family. Dedicated to the world peace.

Acknowledgements

First and foremost, I would like to acknowledge my supervisors Prof. Xiangjian He and Dr. Wenjing Jia for their continuous and endless supervisions and encouragements.

I would like to thank my lab mates, as well as the Dr. Qingqing Wang, Dr. Yue Xi, Dr. Lei Liu, Dr. Saeed Amirgholipour, Dr Mohammadhesam Hesamian, Zhanzhong Gu, Yuanfang Zhang and Chengpei Xu for their assistance or/and encouragements.

I also would like to appreciate the research funding supports from Byker Digital Biotechnology Co. Ltd

Finally, I am particularly grateful to all people who directly or indirectly provided the supports on my research and technical issues, including the netizens in GitHub, and the authors of the papers that I cited.

Ye Huang
Sydney, Australia, 2022.

List of Publications

Journal Papers

1. **Ye Huang**, Qingqing Wang, Wenjing Jia and Xiangjian He, See More Than Once–Kernel-Sharing Atrous Convolution for Semantic Segmentation, in Neurocomputing, Volume 443, 5 July 2021, Pages 26-34.

2. Qingqing Wang, **Ye Huang**, Wenjing Jia, Xiangjian He, Michael Blumenstein, Shujing Lyu and Yue Lu, FACLSTM: ConvLSTM with focused attention for scene text recognition, Science China Information Sciences 2020

Conference Papers

1. **Ye Huang**, Di Kang, Liang Chen, Xuefei Zhe, Wenjing Jia, Xiangjian He, Linchao Bao, CAR: Class-aware Regularizations for Semantic Segmentation, accepted by ECCV 2022.

2. **Ye Huang**, Di Kang, Wenjing Jia, Xiangjian He and Liu Liu, Channelized Axial Attention – Considering Channel Relation within Spatial Attention for Semantic Segmentation, Proceedings of the AAAI Conference on Artificial Intelligence, 36(1), 1016-1025.

3. Qingqing Wang, Wenjing Jia, Xiangjian He, Yue Lu, Michael Blumenstein, **Ye Huang** and Shujing Lyu, DeepText: Detecting text from the wild with multi-ASPP-assembled deeplab, Proceedings of 2019 International Conference on Document Analysis and Recognition

Contents

Abstract iii

Dedication v

Acknowledgments vi

List of Publications vii

List of Figures xii

List of Tables xvii

Abbreviation xxi

1 Introduction 1

1.1 Research Topics in Semantic Segmentation 2

1.1.1 Context Aggregation 2

1.1.2 High Resolution and Details 3

1.2 Early Context Aggregation based Approaches (prior to 2016) 4

1.3 Fixed Range Multi-scale Context Aggregation (2017-2018) 4

1.4 Context Aggregation with Attention Mechanism (since 2018) 7

1.5 Class-aware Context Aggregation (since 2020) 9

1.6 Latest Research Status (2022) 10

1.7 Research Problems and Contributions 11

2 Kernel-Sharing Atrous Convolution 14

2.1 Related Work and Issues 16

2.1.1 Fully Convolutional Network 16

2.1.2 DeepLab Family 17

2.1.3 Other Semantic Segmentation Models 19

2.2 Proposed Solution 20

2.2.1 Atrous Spatial Pyramid Pooling 20

2.2.2 Atrous Convolution with Shared Kernel 21

2.3 Experiment Setting 24

2.3.1 Datasets and Data Augmentation 24

2.3.2 Implementation Details 26

2.4 Experiment Results 29

2.4.1 Improved mIOU 29

2.4.2 Reduced Computational Cost 30

2.4.3 Capability of Handling Wider Range of Context 30

2.4.4 Improved Speed with Less GPU Memory Usage 32

2.4.5 Experiment Results on ADE20K 32

2.5 Summary 33

3 Channelized Axial Attention 35

3.1 Related Work and issues 38

3.2 Exploring Conflicting Features 40

3.2.1 Visualizing Conflicts 41

3.2.2 Examples of Conflicting Features 42

3.3 Methods 44

3.3.1 Preliminaries 44

3.3.2 Channelized Axial Attention 45

3.4 Experiments 48

3.4.1 Implementation Details 48

3.4.2 Results on PASCAL Context 49

3.4.3 Results on COCO-Stuff 10K 55

3.4.4 Results on Cityscapes 56

3.4.5 Results on COCOStuff-164k 56

3.5 Extra Visualizations 57

3.5.1 COCOStuff-10k 57

3.5.2 PASCAL Context 58

3.6 Pseudo Code of Group Vectorization 58

3.7 Limitation 59

3.8 Summary 60

4 CAR: Class-aware Regularization 65

4.1 Introduction 65

4.2 Related Work 68

4.2.1 Self-Attention 68

4.2.2 Class Center 69

4.2.3 Inter-Class Reasoning 69

4.3 Methodology 71

4.3.1 Extracting Class Centers from Ground Truth 71

4.3.2 Reducing Intra-Class Feature Variance 71

4.3.3 Maximizing Inter-class Separation 73

4.3.4 Differences with OCR, ACFNet and CPNet 76

4.4 Experiments 77

4.4.1 Implementation 77

4.4.2 Experiments on Pascal Context 77

4.4.3 Experiments on COCOStuff-10K 84

4.5 Extra Visualizations 86

4.5.1 Visualization of OCRNet on Pascal Context 86

4.5.2 Visualization of DeepLab on Pascal Context 87

4.6 Extra Technical Details 88

4.6.1 Deterministic 88

4.7 Limitation 89

4.8 Summary and Future Work 90

5 Summary and Discussion 93

5.1 Conclusion 93

5.2 Future Work 95

5.2.1 Mining Inter-class Relations with Inter-class Dependent Encodings 95

5.2.2 Decouple Upsampling Encoding 96

Bibliography 97

List of Figures

1.1 Examples of Semantic Segmentation results on COCOStuff [1] dataset 1

1.2 FCN [37] based on VGG outputs pixel level predictions 5

1.3 The design of PSPNet [69], which uses PPM that contains multiple pooling operations with different size rates to aggregate the multi-scale context. 5

1.4 The design of DeepLab V3 [4], which uses ASPP containing multiple dilated (atrous) convolution of different dilation rates to aggregate multi-scale context. 6

1.5 The designs of Self-Attention [53]. It compares the dot-product similarities between each of 2 pixels on the feature map, and then uses it as the weights to aggregate context from all pixels on the feature map. Images come from [72] 8

2.1 The multi-branch-like solutions used in PSPNet and DeepLab for improving models' robustness to objects' scale variability. 16

2.2 Illustration of our proposed Kernel-Sharing Atrous Convolution structure. The single 3 × 3 kernel is shared by three parallel branches with different atrous rates. 17

2.3 The detailed architecture of our proposed Kernel-Sharing Atrous Convolution with $rate = (6, 12, 18)$ 21

2.4 Visualization of the feature maps extracted by kernels of KSAC and ASPP. Here, 25 feature maps are presented for each rate, and we enlarge the ones indicated by red bounding boxes on the top of the figure. Apparently, edges and contours extracted by the shared kernel of our KSAC are much clearer than those extracted by separate kernels of ASPP, for both large atrous rate and small atrous rate. Readers are suggested to zoom in to see more details. . . 23

2.5 Comparison of the segmentation results obtained by FCN, ASPP and our KSAC on the Pascal VOC 2012 validation set. 27

3.1 Different dual attention designs: (a) **Parallel dual attention** sums the results from spatial and channel attentions directly, which may cause conflicts because spatial and channel attentions are focusing on different aspects. (b) **Sequential dual attention** performs spatial attention after channel attention, where the spatial attention may override correct features extracted by the channel attention. (c) **Our channelized attention** seamlessly merges the spatial and channel attentions into a single operation (see Sect. 3.3.2), removing the potential conflicting issue caused by **a** or **b**. 36

3.2 Our designs for visualizing the effects of dual attentions in parallel and sequential. 40

3.3 Conflicting features in parallel dual attention designs. **Top:** The bad spatial attention representation negatively influences the good channel attention representation. **Bottom:** The bad channel attention representation negatively influences the good spatial attention representation. See the boxed areas. 41

3.4 In sequential dual attention designs, the spatial attention representation (the 4th column) ignores the correct channel attention representation (the 3rd column). 41

3.5 The detailed architecture of the proposed CAA. We present the way to apply channel attention seamlessly in (**b**). We mark the independent spatial dimension in the **bold** style. This allows channel attention to also consider spatially unique information. *Note that*, in our design, the "*value*" for Row attention is obtained from the result of Column attention. See Eq. 3.11 for details. 43

3.6 Examples of the segmentation results obtained on the PASCAL Context dataset using FCN, DANet and CAA. 55

3.7 Examples of the results obtained on the COCO-Stuff 10K dataset with our proposed CAA in comparison to the results obtained with FCN, DANet and the ground truth. 62

3.8 Examples of the results obtained on the PASCAL Context dataset with our proposed CAA in comparison to the results obtained with FCN, DANet and the ground truth. 63

3.9 Extra examples of the segmentation results obtained on the Cityscapes validation set [38] with our proposed CAA in comparison to the results obtained with DANet [14] and the ground truth. 64

4.1 The concept of the proposed CAR. Our CAR optimizes existing models with three regularization targets: 1) reducing pixels' intra-class distance, 2) reducing inter-class center-to-center dependency, and 3) reducing pixels' inter-class dependency. As highlighted in this example (indicated with a red dot in the image), with our CAR, the grass class does not affect the classification of dog/sheep as much as before, and hence successfully avoids previous (w/o CAR) mis-classification. 67

4.2 The difference between the proposed CAR and previous methods that use class-level information. Previous models focus on extracting class center while using simple concatenation of the original pixel feature and the class/context feature for later classification. In contrast, our CAR uses direct supervision related to class center as regularization during training, resulting in small intra-class variance and low inter-class dependency. See Fig. 4.1 and Sect. 4.3 for details. 70

4.3 **The proposed CAR approach.** CAR can be inserted into various segmentation models, right before the logit prediction module (A1-A4). CAR contains three regularization terms, including (C) intra-class center-to-center loss $\mathcal{L}_{\text{intra-c2p}}$ (Sect. 4.3.2.2), (D) inter-class center-to-center loss $\mathcal{L}_{\text{inter-c2c}}$ (Sect. 4.3.3.2), and (E) inter-class center-to-pixel loss $\mathcal{L}_{\text{inter-c2p}}$ (Sect. 4.3.3.3). 72

4.4 Visualization of the feature similarity between a given pixel (marked with a red dot in the image) and all pixels, as well as the segmentation results on Pascal Context test set. A hotter color denotes larger similarity value. Apparently, our CAR reduces the inter-class dependency and exhibits better generalization ability, where energies are better restrained in the intra-class pixels. 85

4.5 Class dependency maps generated on Pascal Context test set. One may zoom in to see class names. A hotter color means that the class has higher dependency to the corresponding class, and vice versa. It is obvious that our CAR reduces the inter-class dependency, thus providing better generalizability (see Figs. 4.1 and 4.4). 86

4.6 Visualization of the feature similarity between a given pixel (marked with a red dot in the image) and all other pixels, as well as the segmentation results of **HRNetW48 [52] + OCR [61]** on Pascal Context test set. A hotter color denotes a greater similarity value. . 91

4.7 Visualization of the feature similarity between a given pixel (marked with a red dot in the image) and all pixels, as well as the segmentation results of **ResNet-50** [**18**] + **DeepLab** [**4**] on Pascal Context test set. A hotter color denotes a greater similarity value. 92

List of Tables

2.1 Experimental results obtained on PASCAL VOC 2012 validation set with different inference strategies when using ASPP and our proposed KSAC, and ResNet-50, ResNet-101, Xception65 or MobileNetV2 as the backbone. **KSAC:** Using our proposed KSAC. **ASPP:** Using the standard ASPP structure proposed in DeepLabv3+ [4]. **D:** Concatenating the OS = 4 feature maps from the backbone during the upsampling of the logits. **MF:** Employing multi-scale (MS) and left-right flipping on the inputs during the evaluation. **COCO:** Model is pre-trained on COCO dataset [1]. The Performance is evaluated from the aspect of **mIOU (%)** and the number of FLOPS (Floating Point Operations per Second), respectively. 28

2.2 Experimental results of our proposed KSAC on PASCAL VOC 2012 validation set with different settings of atrous rates. 31

2.3 Comparison results with other approaches on the PASCAL VOC 2012 validation and test sets. 31

2.4 Comparison results with other approaches on the ADE20K validation set for multi-scale prediction. 33

3.1 Results without using channelization (Row 1) and using channelization with different layer counts and channel numbers. Numbers in parentheses indicate standard deviations (see Sect. 3.4.2). 50

3.2 Result comparison between axial attention, axial attention + SE and channelized axial attention. 51

3.3 Ablation study of applying our Channelized Attention on self-attention with ResNet-101. **Eval OS**: Output strides [5] during evaluation. 51

3.4 Comparing results with different testing strategies. **OS:** Output stride in training and inference. **MF:** Apply multi-scale and left-right flipping during inference. **Aux:** Add auxiliary loss during training. "+" refers to the extra FLOPS over the baseline FLOPS of ResNet-101. 52

3.5 Ablation study of CAA with the backbones other than ResNet-101. All results are obtained in single scale without flipping. **OS**: Output strides during evaluation. **AA**: Axial Attention. **C**: Channelized. . . . 53

3.6 Comparisons with other state-of-the-art approaches on the PASCAL Context test set. For a fair comparison, all compared methods use ResNet-101 and naive upsampling. 54

3.7 Result comparison with the state-of-the-art approaches on the PASCAL Context test set for multi-scale prediction. Note that, the listed methods were not trained under the same settings, or using same backbone. 56

3.8 Comparisons with other state-of-the-art approaches on the COCO-Stuff 10K test set. For a fair comparison, all compared methods use ResNet-101 and naive upsampling. 57

3.9 Result comparison with the state-of-the-art approaches on the COCO-Stuff-10K test set for multi-scale prediction. Note that, the listed methods were not trained under the same settings, or using same backbone. 58

3.10 Comparisons with other state-of-the-art approaches on the Cityscapes test set. For a fair comparison, all compared methods use ResNet-101 and naive upsampling. 59

3.11 Result comparison with the state-of-the-art approaches on the COCO-Stuff-164K test set for multi-scale prediction. Note that, the listed methods were not trained under same settings, or using same backbone. *Methods* other than CAA and Segformer are reproduced in Segformer paper. 60

4.1 Ablation studies of adding CAR to different methods on Pascal Context dataset. All results are obtained with single scale test without flipping. “A” means replacing the 3×3 conv with 1×1 conv. CAR improves the performance of different types of backbones (CNN & Transformer) and head blocks (SA & Uper), showing that the proposed CAR generalizes well on different network architectures. 78

4.2 Ablation studies of adding moving average to CAR on Pascal Context. The decay rate stands for the effect of old class center. . . . 80

4.3 Comparison of mIOUs (%) obtained when using the batch class center vs using the image class center in CAR. 80

4.4 The computational cost (in GFLOPs) of the proposed CAR on a 513×513 image with an output stride of 8. 81

4.5 Ablation studies of adding CAR to different baselines on Pascal Context [41] and COCOStuff-10K [1]. We deterministically reproduced all the baselines with the same settings. All results are obtained with single-scale testing without flipping. CAR works very well in most existing methods. **§** means reducing the class-level threshold ϵ_0 from 0.5 to 0.25. We found it is sensitive for some model variants to handle a large number of class. Affinity loss [60] and Auxiliary loss [69] are applied on CPNet and OCR, respectively, since they highly rely on those losses. 83

4.6 Experiments on boosting the SOTA single-model performance on Pascal Context by our CAR. See Sect. 4.4.2.9 for the details. *§*: We report previous SOTA scores as reference. *SS*: Single scale without flipping. *MF*: Multi-scale with flipping. JPU is used to get features with output stride = 8. *Aux*: Apply auxiliary loss during training (see [69]). *Iterations*: training iterations. 87

4.7 Experiments on boosting SOTA on COCOStuff10k, levering the previous single model SOTA and boosted by our CAR. See Sect. 4.4.3.2 for details. *§*: We report the original SOTA scores. *SS*: Single scale without flipping. *MF*: Multi-scale with flipping. *Aux* Apply auxiliary loss during training, see [69]. 88

4.8 Ablation studies of our proposed CAR using different random seeds on the Pascal Context dataset. 89

Abbreviation

- H - Height
- W - Width
- Channels - The size of last dimension of the 4D feature map.
- Encoding - Nerual network encoded 3 channels image input to the faeture map with multiple channels.
- mIOU - Mean Intersection over Union
- KSAC - Kernel Sharing Astrous Convolution
- CAA - Channelized Axial Attention
- CAR - Class-aware Regularization
- OS - Output stride [3]
- FCN - Fully Convolutional Networks [37]
- ASPP - Atrous Spatial Pyramid Pooling [4]
- PPM - Pyramid Pooling module [69]
- FPN - Feature Pyramid Networks [32]
- SA - Self-attention [53]
- ACFNet - Attentional Class Feature Network [64]
- OCR - Object-Contextual Representations [61]
- CPNet - Context Prior Network [60]