Open World Semantic Segmentation

Manikanda Krishnan V

A Thesis Submitted to

Indian Institute of Technology Hyderabad

In Partial Fulfillment of the Requirements for

The Degree of Master of Technology

Department of Computer Science and Engineering

June 22, 2022

Declaration

I declare that this written submission represents my ideas in my own words, and where ideas or words of others have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be a cause for disciplinary action by the Institute and can also evoke penal action from the sources that have thus not been properly cited, or from whom proper permission has not been taken when needed.

(Signature)

Manikanda Krishnan V

(Name)

CS19MDS11033

(Roll No.)

Approval Sheet

This Thesis entitled **Open World Semantic Segmentation** by Manikanda Krishnan V is approved for the degree of Master of Technology from IIT Hyderabad.

_____

(Dr. Vineeth N Balasubramanian) Adviser
Dept. of Computer Science and Engineering
Indian Institute of Technology Hyderabad

Acknowledgements

I would like to express my gratitude to my supervisor Prof. Dr Vineeth.N.Balasubramanian and his PhD student Joseph.K.J for introducing me to this exciting problem statement and providing me the guidance needed to work on this project.

Abstract

Convolutional Neural Networks have been pivotal in the domain of semantic segmentation. The traditional problem statement assumes that the classes that are encountered during inference are a subset of the classes that have been shown to the network during the training phase i.e a closed set of classes. This assumption is usually violated in real world computer vision applications where there is little control over the kind of inputs the model would be subject to i.e unseen classes. This requires the segmentation model to first identify unseen objects in the image and then incrementally learn to segment these unknown classes.

This work proposes to address this challenge via a two step approach. In step 1, we train a prototypical segmentation network that performs close-set segmentation. The class scores are computed via cosine similarity. The anomalous regions are then detected via max-softmax function over the pixel values. The novel objects are then learned via finetuning a new prototype to align with the embeddding of the novel class.

Contents

List of Tables

List of Figures

List of Algorithms

Chapter 1

Introduction

1.1 Motivation

Deep Learning has been tremendously successful in the task of segmenting complex scenes by exploiting information from large high quality datasets. However most of these explorations have been in situations where the classes to be segmented during inference are a subset of the classes seen during the training phase i.e close-set and static situations. Close-set systems could cause potential harm especially in safety related applications - if the model incorrectly segments an unknown object as a known class. Even in other applications - a system that is able to say ”don’t know” when faced with the unknown is better than a system that makes incorrect guesses. Once the unknown object is identified the system can be trained further to learn to segment them. This iterative cycle consisting of closed set segmentation -> anomaly detection -> incremental learning constitutes the Open World Segmentation paradigm.

This thesis aims to study and implement an approach that can segment semantic classes reasonably well in an open world setting.

1.2 Contributions

1. This thesis provides a simplified approach to the task of Open World Detection.
2. An orthogonalization term is added to learn better feature representations.
3. A simple but viable approach is used for incremental segmentation

1.3 Thesis Outline

The thesis presents an overview of Open World Segmentation and its important components such as uncertainty quantification and incremental segmentation. It is organized as follows

1. *Chapter 2* presents a review of associated literature. It covers the topics related to semantic segmentation, uncertainty quantification, incremental segmentation and existing work on open world semantic segmentation.

2. *Chapter 3* describes the approach in detail and explains the motivation behind why certain decisions were made.

3. *Chapter 4* describes the data and experiment details so that the work can be reproduced by those who are interested.

4. *Chapter 5* presents the results of the experiments in both quantitative and qualitative manner and draws conclusions from them.

Chapter 2

Related Work

Open world semantic segmentation is built on top of semantic segmentation, uncertainty quantification and incremental segmentation algorithms. This chapter aims to describe and categorize the existing literature available on the above.

2.1 Semantic Segmentation

Semantic segmentation relies on the quality of high level features. These high level features generally occur in deeper layers of the network[1] due to the high receptive fields captured by each neuron here. The cost however is the reduction in resolution of the resulting feature maps which results in subpar results. Deep learning based segmentation algorithms can hence be characterized by the resolution at which the segmentation activity takes place.

2.1.1 Coarse Resolution Segmentation

Since deeper layers of classification architectures capture high level semantics Fully Convolutional Network(FCN)[2] like networks replace the global pooling/flattening layers with a $1 * 1$ convolution to perform pixel wise classification. The resultant coarse segmentation map is then converted to image resolution by upsampling via bilinear interpolation or strided convolutions.

Dilated Convolutions[3] were introduced to capture larger receptive fields and hence more complex features at higher resolutions. This results in an overall improvement to the resolution of the generated segmentation maps especially in the segmentation of smaller objects.

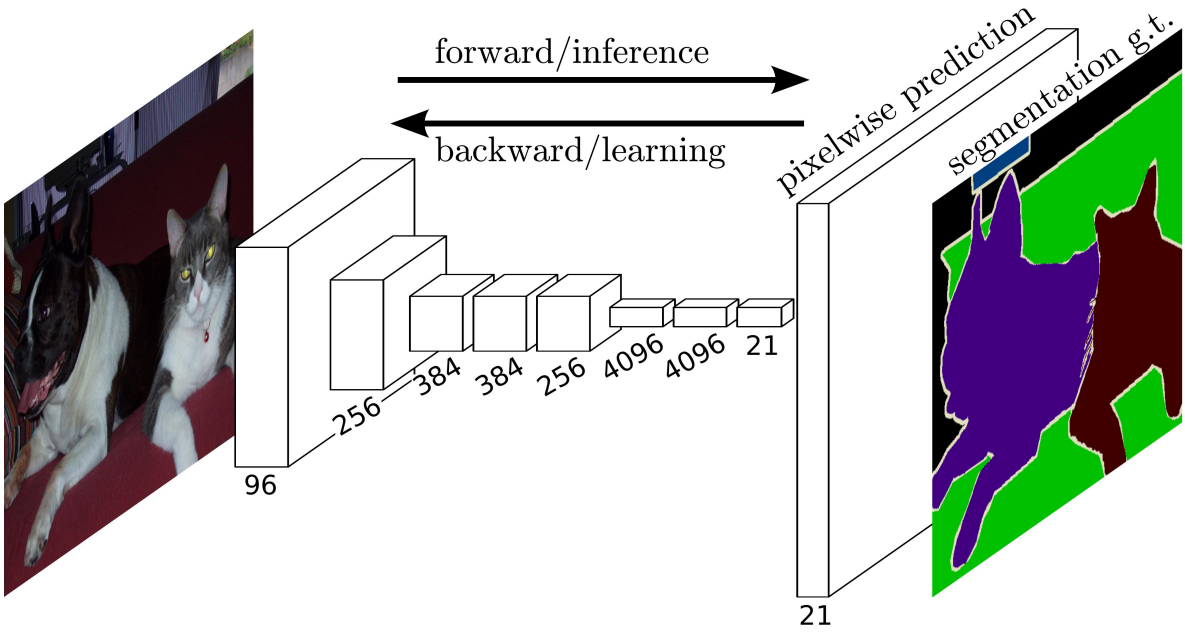

Figure 2.1: Fully Convolutional Network

2.1.2 Coarse To Fine Resolution Segmentation

UNet[4] introduced skip connections from the downsampling netowrk to the upsampling layers, thereby incorporating low-level features to enhance the detail of the full-resolution segmentation map. The skip connections have the additional advantage of easing backpropagation through the downsampling network. Feature Pyramid Network(FPN)[5] like networks add detection/segmentation heads to the intermediate upsampling layers to detect/segment objects at scale. These basic frameworks have been extended with modifications to the down and up sampling backbones [5], interactions between skip connections[6] etc to make the transfer of information from the downsampling stage to the upsampling stage more efficient.

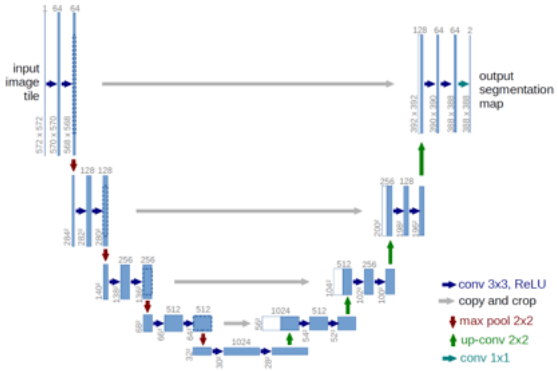

Figure 2.2: U-Net

2.1.3 Preserving High Resolution Information

Convolutional Neural Fabrics[7], UNet++[8], High Resolution Network(HRNet)[9] builds on the findings from all the above to extract high level features at multiple resolutions making it extremely useful in applications where objects occur at various scales. These networks generally have multiple pathways with each pathway processing a feature map at different resolutions. At the end of every stage the feature maps from other pathways are combined to provide comprehensive information to the next stage of the network.

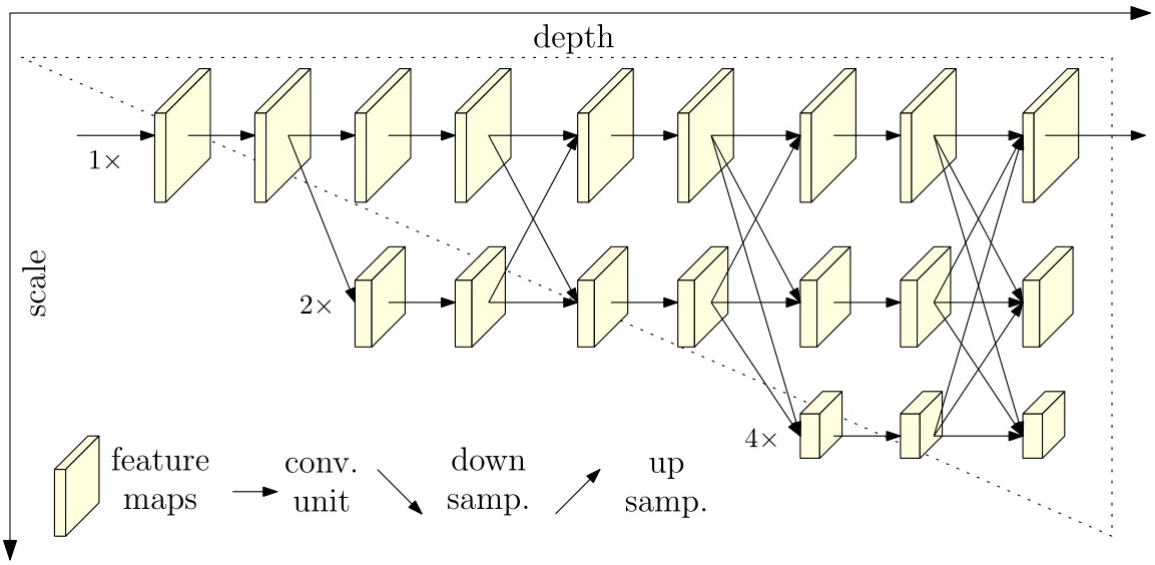

Figure 2.3: High Resolution Network(HRNet)

2.2 Uncertainty Quantification in Deep Learning

An intelligent system must be capable of stating when it doesn't know the output for given set of inputs. This would improve the explainability of the model and make it more viable for actual use. This is especially important in problems where a wrong classification could lead to very undesirable results such as in self driving cars.

2.2.1 Types of Uncertainty

There are two main types of uncertainty in bayesian modelling[10].

1. **Aleatoric Uncertainty**: This uncertainty results from the underlying stochasticity of the input. This uncertainty cannot be reduced with increase in data.

2. **Epistemic Uncertainty**: This uncertainty arises from limitations of the learned model. These limitations might arise due to biases and availability of data. This uncertainty can be reduced by adding the relevant data to the training phase.

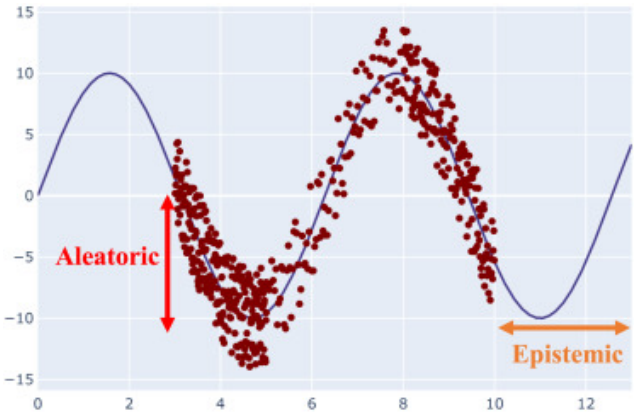

Figure 2.4: Types of Uncertainty

2.2.2 Methods for Epistemic Uncertainty Estimation

1. **Monte Carlo(MC) Dropout**: [11] showed us that networks with dropouts in every layer approximates a deep gaussian process. Using dropouts during inference can hence generate MC samples which can then be used to approximate probability distributions over classes. Dropout networks are generally robust with respect to the output for in-distribution samples due to the nature of training procedure. Hence samples with high estimated entropy are more likely to belong to out-of-distirbution(OOD) samples. Bayesian SegNet [12] is an example of this technique applied to a segmentation task.

2. **Deep Ensembles**: [13] showed that good estimates for uncertainty can be obtained by taking an ensemble of networks initialized with different weights.

3. **Evidential Regression**: In this class of methods[14] parameters to a prior distribution from which the mean and variance of an output variable is sampled from. OOD Samples are expected to have a larger variance in mean values.

4. **Image Resynthesis**: In this class of methods[15] the image is resynthesized. These methods are founded on the premise that the reconstruction for OOD samples would be poorer than the reconstruction for in distribution samples.

5. **Maximum Logit**: [16] shows us that the maximum logit can be used as a rough estimate of foreground probability especially in the problem of multi label classification.

6. **Deep Metric Learning** In deep metric learning[17] the network is trained to learn feature representations in some learned metric space. This is achieved by training to minimize/maximize the distance computed via some distance metric such as Mahalonobis, Manhattan, Euclidean etc between the learned embedding of similar/dissimilar samples. They can be combined with learned prototypes [18] to estimate the likelihood of a sample belonging to one of the in-distribution classes.

2.3 Few Shot Segmentation

There are many situations where a comprehensive dataset covering all scenarios can be collected for model training. Intelligent agents deployed in such scenarios must have the capability to learn new semantic information preferably from few samples to be useful. Due to catastrophic forgetting algorithms special algorithms need to be developed to ensure that learned information is not lost when fed with new data [19]. Few shot semantic segmentation algorithms can be classified into 2 major approaches

1. **Prototypical Networks**: In these methods[20, 21, 22, 23, 24] the class label is assigned based on the distance between the learned pixel representation and the prototype. The prototype for a new class is usually computed by Masked Average Pooling(MAP) which averages the learned representation of the pixels corresponding to that class. Some of these approaches extend this to predict multiple prototypes per class.

2. **Feature Combination**: In these methods [25, 26, 27, 28, 29] the support set and query image features are combined in some manner usually at the end of the encoder stage to jointly decode the common class.

2.4 Open World Semantic Segmentation

Open world semantic segmentation problem extends the semantic segmentation problem to targets that contain unseen semantic categories. This extension necessitates the detection and subsequent segmentation of out-of-distribution classes.

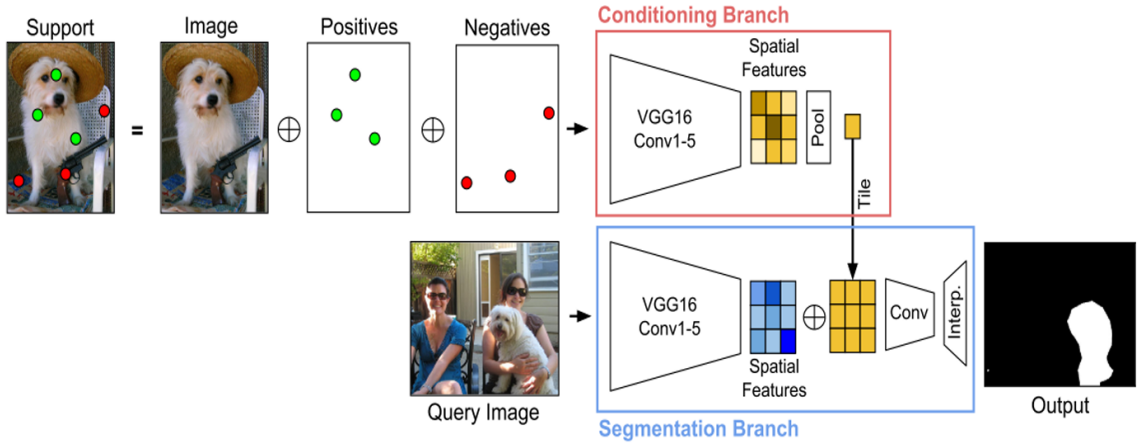

Figure 2.5: co-FCN an example of feature concatenation based Few-Shot Segmentation Algorithm

Deep Metric Learning for Open World Semantic Segmentation[30] proposes to address this problem by dividing it into two stages. The first stage involves training a metric learning network to cluster the samples around prototypes based on euclidean distance metric and the second stage involves incremental few shot segmentation via a prototype based algorithm. Anomalous classes are detected via computing a energy function based on the distance between the predicted embedding and prototypes.

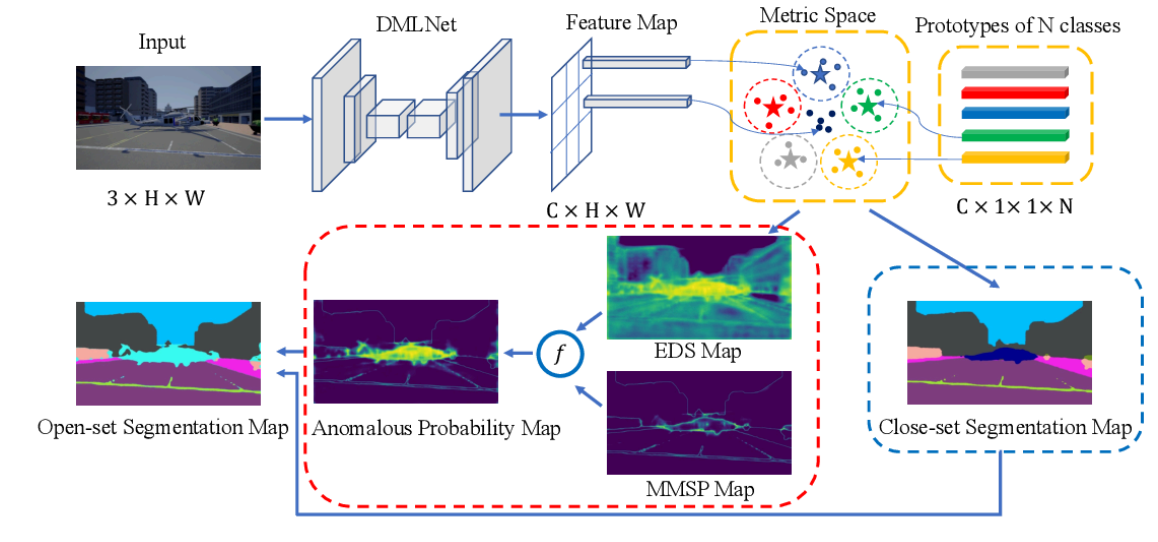

Figure 2.6: Deep Metric Learning for Open World Semantic Segmentation

Region-Aware Metric Learning for Open World Semantic Segmentation via Meta-Channel Aggregation[31] use additional pseudo-output classes during the closed set segmentation phase. These additional channels would act as buffers where the un-

known classes would aggregate during the training process thereby greatly improving the anomaly detection step. This aggregation is made robust by imposing non overlapping constraints which they claim makes the network learn segment parts of unknown objects. The final segmentation for novel classes is done by combining multiple meta-channels to generate the novel class segmentation

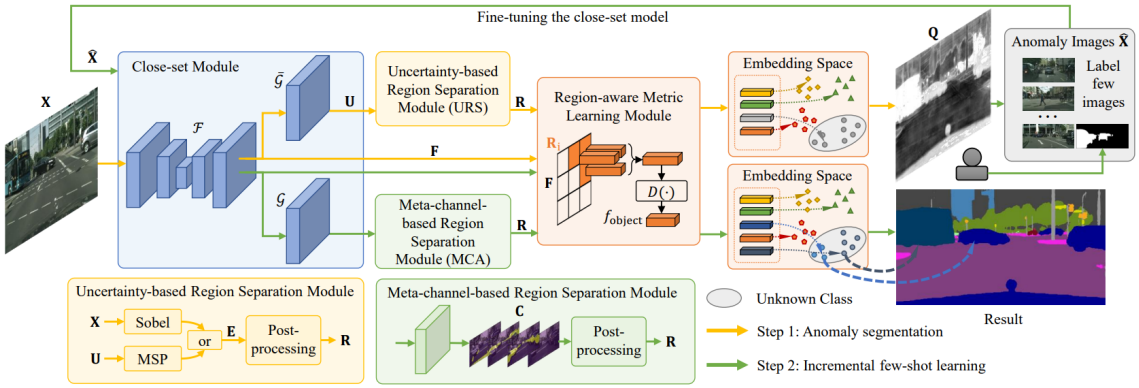

Figure 2.7: Region Aware Metric Learning for Open World Semantic Segmentation via Meta-Channel Aggregation

Known Region Aware Domain Alignment (KRADA)[32] unlike the above two papers feeds both source and target images during the training phase. It is more concerned with applying the segmentation model to a setting with a distribution shift in the input domain which they solve by feature alignment using Generative Adversarial Networks(GAN). The unknown classes which shouldn't be aligned are masked out from the feature alignement step via a pseudo label generation mechanism for the target images.

The segmentation performance of novel classes in the above approaches steeply declines when the number of novel classes is greater than one.

2.4.1 Summary

This chapter provides a high level overview of the techniques that are involved in the components that make Open World Semantic Segmentation(OWSS) possible. The existing methods on OWSS suffer significant performance drops when the novel classes are more than one. This work hopes to address this challenge.

Chapter 3

Methodology

This chapter explains the approach3.1 that has been undertaken.

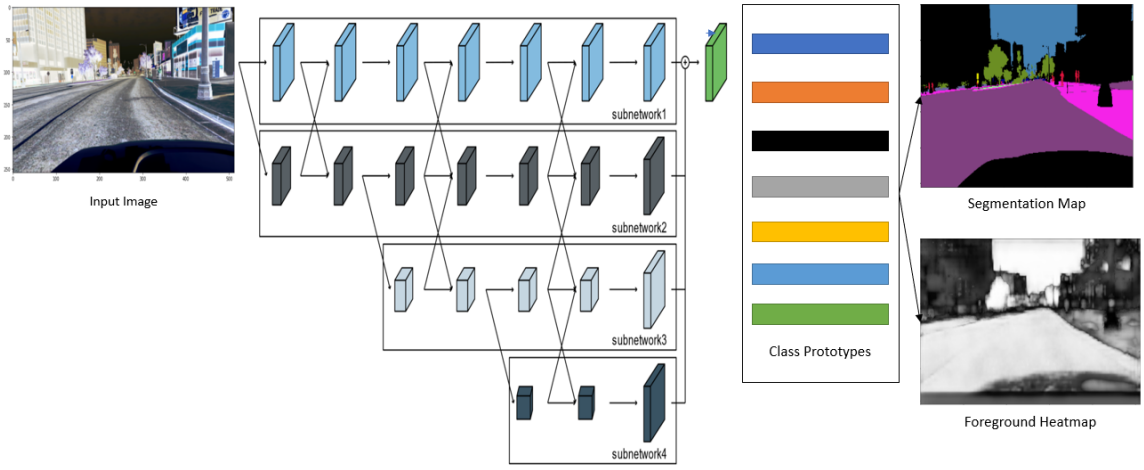

Figure 3.1: The image is fed into the HRNet segmentation model. The extracted features are then compared with the class prototypes using cosine similarity to generate the scores. The segmentation labels are assigned based on maximum similarity and the foreground heatmap is generated via MSP

3.1 Network Architecture

HighResolution Network is used as a feature extractor. The strides of the first 2 convolutional layers are changed from 2 to 1 to make the network operate at original input resolution. Batch Normalization is replaced with Instance Normalization + 2D Dropout. The multi-resolution feature outputs are resampled to generate the full resolution feature map $F_{img}, ||F_{img}||_2 = 1$. The network has 2 output branches which are listed below:

1. **Segmentation Head**: The segmentation head consists of $C^{input} + C^{meta}$ unit normed prototypes ${w^{seg}}_c, ||w^s eg_c||_2 = 1$. The C^{meta} classes are additional channels that serve as buffers for the background objects. The cosine similarity is computed between the input and output to get the output vector O_{cls}. Since O_{cls} represents the cosine similarity between the feature vector and prototype vector all its elements are in the range $[0, 1]$.

2. **Image Reconstruction Head**: F_{img} is then passed through $1 * 1$ convolutions to reconstruct the RGB image.The purpose of the reconstruction task is to teach the network to reason about the scene. To achieve this some regions of the images are cutout and the network is made to reconstruct the actual image from this incomplete version.

The weight vectors in the segmentation head are encouraged to be orthogonal by adding an additional term to the loss function. This is done to encourage the network to learn nearly orthogonal feature representations of the semantic classes.

3.2 Anomaly Detection

We can use cosine similarity scores to define a conditional likelihood. This can then be used to frame the objective as a likelihood maximization problem.

Let $C = \{C_1, C_2, ...C_N\} \& P(C_i) = P(C_j) \forall i, j$

Let $P(\frac{C = c}{Embedding = e}) = \frac{| < e, w_c^{seg} > |}{\Sigma_{i \in C} | < e, w_i^{seg} > |}$

This is then used to define the Maximum Softmax Probability to detect the foreground objects.

3.3 Incremental Segmentation

The feature extraction network when trained well would have already captured relevant features from these novel objects. All that remains is to learn a unit vector that is more aligned with the embedding of this novel class. We achieve this by freezing all weights and finetuning only the new weight vector. This is similar to the work described in [33]

Chapter 4

Data and Experimental Setup

This chapter explains our experimental setup in detail to facilitate the reproduction of this work by all who are interested.

4.1 Data

4.1.1 CityScapes Dataset

Cityscapes[34] is a dataset created for urban scene understanding. It contains road scenes as shown in Fig 4.1 from 50 different cities out of which 19 are set aside for training. The semantic categories in this dataset are depicted in Fig 4.2. There are around 2000 images in the training set and 500 images part of the evaluation set.

The images are downsampled to 1024x512 for training the models. Images containing Bus, Truck, Train, Motorcycle are removed from the training set.

4.1.2 GTA5 dataset

GTA5 dataset introduced in [35] contains 24000 synthetic images for urban scene understanding obtained from video games. The labels are compatible with the cityscape dataset. Roughly half of the images are used for training. Images containing Bus, Truck, Train, Motorcycle are removed. The test set contains 6181 images.

4.1.3 Fishyscape/Lost&Found dataset

Fishyscapes/Lost&Found [36, 37] is an anomaly detection dataset containing road scenes with anomalous objects like animals, litter etc lying on the roads as shown in

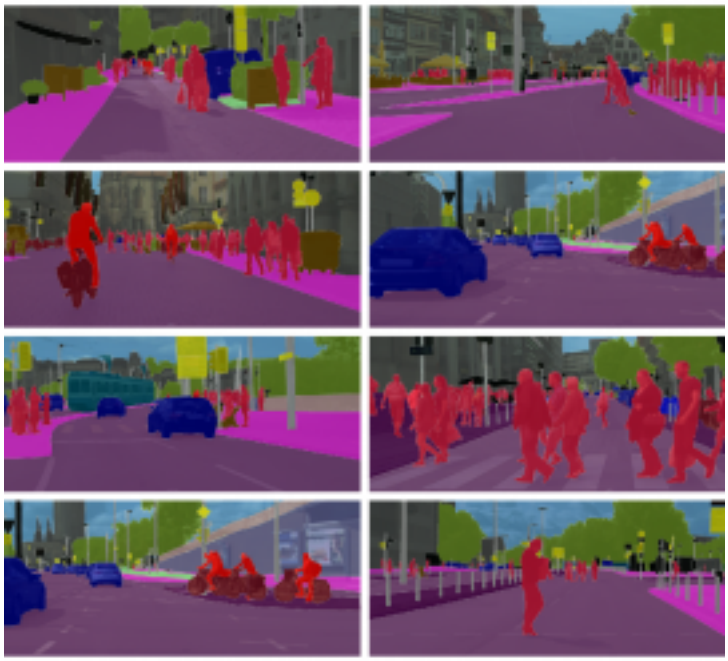

Figure 4.1: Examples from cityscapes dataset

Figure 4.2: Semantic Labels present in cityscapes dataset

Fig 4.4. This contains around 100 samples and all of this is used for evaluating the model that has been trained on the above two datasets.

4.2 Experiment

All models were initialized with the same seed.

4.2.1 Baseline

A standard segmentation training pipeline was run to train HRNet. There was no orthogonalization constraints used during the training phase.

4.2.2 Data Augmentations

The images are augmented by applying color-jitters, random posterization, inversion, solarization, sharpness and contrast adjustments.

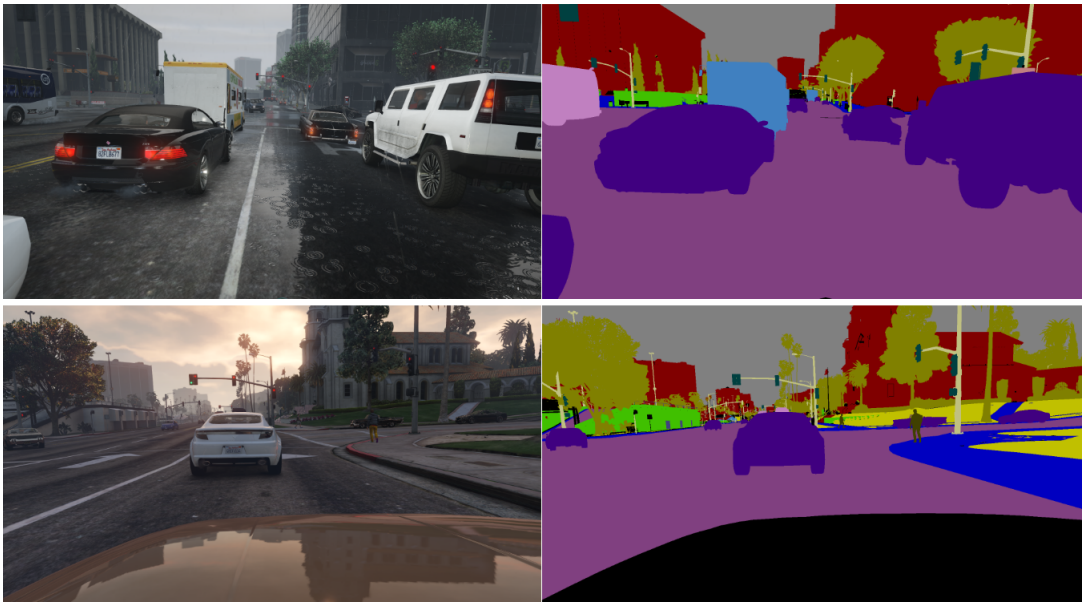

Figure 4.3: GTA5 dataset

Figure 4.4: Lost&Found dataset

4.2.3 Loss Functions

The cosine similarity based conditional likelihood was maxmized via cross entropy and soft-dice loss.

The reconstruction error was minimized using Mean Absolute Error.

Orthogonalization was encouraged by using L2 loss on the dot products between the prototypes.

4.2.4 Training Parameters

Adam Optimizer with a learning rate of $1e^{-2}$ was used for training the network. A scheduler with exponential decay(value=0.9) was used for adjusting the learning after every epoch. The entire model was run for around 20 epochs. In each epoch 6000 samples were sampled uniformly across all classes.

4.2.5 Incremental Segmentation

For training novel classes, 5 images were chosen at random for each novel class. Additional prototypes were added and the network was finetuned for 1000 epochs to learn the prototype orientation while all other parameters where frozen.

4.2.6 Evaluation Metrics

1. The segmentation performance was measured using mean Intersection over Union.

2. The anomaly detection performance was measured using AUROC, FPR@95. As discussed in [16] the scores are computed for every image and then averaged across the samples.

Chapter 5

Results and Conclusions

5.1 Results

The overall results have been tabulated in Table 5.1 and Table 5.2.

5.1.1 Close-Set Segmentation

From Table 5.1 and Table 5.2 it can be seen that the use of orthogonalization constraints enabled the network to locate bridges and cycles. But it can also be observed that there is a 4-5% drop in overall performance when trained with this additional term. Some qualitative results are shown in Fig 5.2.

5.1.2 Anomaly Detection

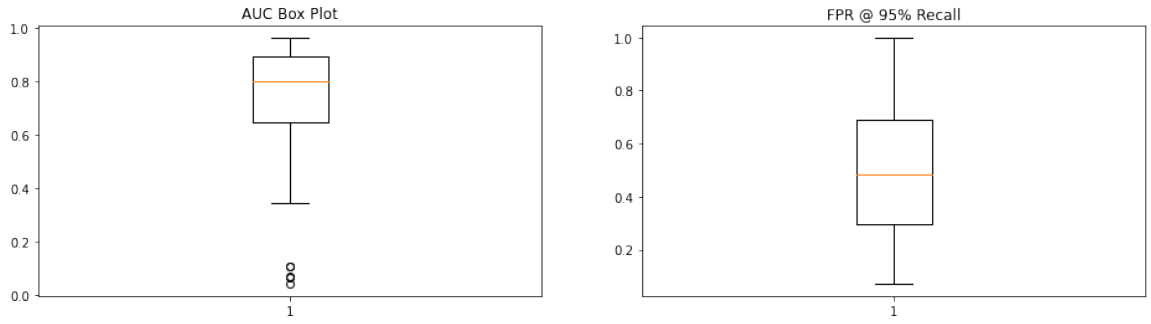

Figure 5.1: AUROC and FPR@95% Recall on Lost and Found Dataset

Fig 5.1 shows the scores across samples of the dataset. The median values for AUROC is around 80% and the median FPR@95% Recall is around 42%. The mean AUROC score is around 74.4% and the mean FPR@95% Recall is around 49%.

Table 5.1: Results on GTA5 Dataset. The columns correspond to the baseline, with the proposed orthogonalization term and the results of incremental segmentation

Type	Baseline	with $\perp$	Inc. Seg
road	84	76	77
sidewalk	64	58	59
parking			
building			
wall	23	16	18
fence	9	5	5
guardrail	12	7	6
bridge	0	5	4
pole			
traffic light	33	19	20
traffic sign	23	10	11
vegetation	71	54	58
terrain	28	16	17
sky	89	74	77
person	32	16	16
rider	14	8	8
car	47	38	46
bicycle	0	27	
novel$_t$*ruck*			
novel$_b$*us*	0	0	1
novel$_t$*rain*	0	0	0
novel$_m$*otorcycle*	0	0	
mIoU_closed	35.27	28.6	30.14
mIoU_novel	0	0	0.5

5.1.3 Incremental Segmentaion

From Table 5.2 it is shown that the proposed scheme improves the detection by 6% on novel classes with a drop of around 3% on the closed set classes in the cityscapes dataset. From Table 5.1 it can be seen that there is no significant gain in the segmentation of novel classes but there is a 2% improvement in the segmentation of the old classes. In Fig 5.2 it can be noticed that there is some confusion between novel class "train" and the existing semantic label "building".

5.2 Discussion and Conclusions

1. The performance drop in the closed-set segmentation problem using the proposed approach may be due to insufficient number of epochs. The normalization

Table 5.2: Results on Cityscapes Dataset. The columns correspond to the baseline, with the proposed orthogonalization term and the results of incremental segmentation

Type	Baseline	with $\perp$	after Inc Seg
road	89	77	77
sidewalk	61	53	54
parking	30	22	22
building	66	66	61
wall	23	18	17
fence	24	20	17
guardrail	37	27	27
bridge	0	15	11
pole	44	31	30
traffic light	38	24	23
traffic sign	47	34	35
vegetation	82	79	80
terrain	35	27	27
sky	70	62	63
person	47	32	30
rider	18	18	12
car	65	67	68
bicycle	0	33	7
novel$_t$*ruck*	0	0	2
novel$_b$*us*	0	0	5
novel$_t$*rain*	0	0	11
novel$_m$*otorcycle*	0	0	8
mIoU_closed	43.11	39.17	36.72
mIoU_novel	0	0	6.5

step would suppress gradients and might have slowed down the optimization process.

2. Even though there is a drop in overall performance, the proposed approach has been able to discover 2 additional classes that were missed in the baseline.

3. The simple finetuning based approach does show the ability to segment novel classes. Even when it fails to offer improvements, it has been able to improve the performance of the closed-set by removing confusion.

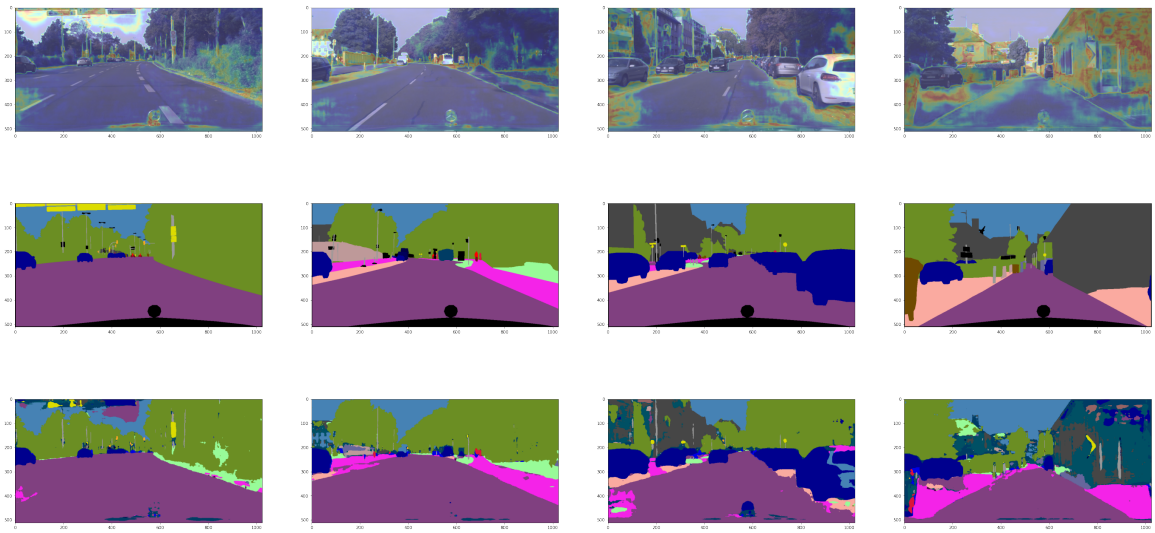

Figure 5.2: Qualitative Results of the proposed segmentation algorithm after incremental learning. The 1^{st} row is the raw images, the 2^{nd} row shows the ground truth and the 3^{rd} row contains the predicted output

References

[1] M. D. Zeiler and R. Fergus. Visualizing and Understanding Convolutional Networks. In Computer Vision – ECCV 2014, 818–833. Springer International Publishing, 2014. 3

[2] E. Shelhamer, J. Long, and T. Darrell. Fully Convolutional Networks for Semantic Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, (2017) 640–651. 3

[3] F. Yu, V. Koltun, and T. Funkhouser. Dilated Residual Networks. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017 . 3

[4] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Lecture Notes in Computer Science, 234–241. Springer International Publishing, 2015. 4

[5] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature Pyramid Networks for Object Detection. In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017 . 4

[6] Z. Zhang, X. Zhang, C. Peng, X. Xue, and J. Sun. ExFuse: Enhancing Feature Fusion for Semantic Segmentation. In Computer Vision – ECCV 2018, 273–288. Springer International Publishing, 2018. 4

[7] S. Saxena and J. Verbeek. Convolutional Neural Fabrics 2016. 5

[8] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang. UNet: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation. *IEEE Transactions on Medical Imaging* 39, (2020) 1856–1867. 5

[9] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao. Deep High-Resolution Representation Learning

for Visual Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, (2021) 3349–3364. 5

[10] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, V. Makarenkov, and S. Nahavandi. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion* 76, (2021) 243–297. 5

[11] Y. Gal and Z. Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In M. F. Balcan and K. Q. Weinberger, eds., Proceedings of The 33rd International Conference on Machine Learning, volume 48 of *Proceedings of Machine Learning Research*. PMLR, New York, New York, USA, 2016 1050–1059. 6

[12] A. Kendall, V. Badrinarayanan, and R. Cipolla. Bayesian SegNet: Model Uncertainty in Deep Convolutional Encoder-Decoder Architectures for Scene Understanding. In Procedings of the British Machine Vision Conference 2017. British Machine Vision Association, 2017 . 6

[13] B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds., Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc., 2017 . 6

[14] A. Amini, W. Schwarting, A. Soleimany, and D. Rus. Deep Evidential Regression . 6

[15] K. Lis, K. K. Nakka, P. Fua, and M. Salzmann. Detecting the Unexpected via Image Resynthesis. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2019 . 6

[16] D. Hendrycks, S. Basart, M. Mazeika, A. Zou, J. Kwon, M. Mostajabi, J. Steinhardt, and D. Song. Scaling Out-of-Distribution Detection for Real-World Settings. *ICML* . 7, 15

[17] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Deep Metric Learning for Person Re-identification. In 2014 22nd International Conference on Pattern Recognition. IEEE, 2014 . 7

[18] H.-M. Yang, X.-Y. Zhang, F. Yin, and C.-L. Liu. Robust Classification with Convolutional Prototype Learning. In 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2018 . 7

[19] E. Belouadah, A. Popescu, and I. Kanellos. A comprehensive study of class incremental learning algorithms for visual tasks. *Neural Networks* 135, (2021) 38–54. 7

[20] G. Li, V. Jampani, L. Sevilla-Lara, D. Sun, J. Kim, and J. Kim. Adaptive Prototype Learning and Allocation for Few-Shot Segmentation. In CVPR. 2021 . 7

[21] L. Yang, W. Zhuo, L. Qi, Y. Shi, and Y. Gao. Mining Latent Classes for Few-shot Segmentation. In ICCV. 2021 . 7

[22] K. Wang, J. H. Liew, Y. Zou, D. Zhou, and J. Feng. PANet: Few-Shot Image Semantic Segmentation With Prototype Alignment. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2019 . 7

[23] B. Yang, C. Liu, B. Li, J. Jiao, and Y. Qixiang. Prototype Mixture Models for Few-shot Semantic Segmentation. In ECCV. 2020 . 7

[24] F. Cermelli, M. Mancini, Y. Xian, Z. Akata, and B. Caputo. Prototype-based Incremental Few-Shot Semantic Segmentation. 2020 . 7

[25] W. Liu, C. Zhang, G. Lin, and F. Liu. CRNet: Cross-Reference Networks for Few-Shot Segmentation. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020 . 7

[26] X. Zhang, Y. Wei, Y. Yang, and T. S. Huang. Sg-one: Similarity guidance network for one-shot semantic segmentation. *IEEE Transactions on Cybernetics* . 7

[27] T. Hu, P. Yang, C. Zhang, G. Yu, Y. Mu, and C. G. M. Snoek. Attention-Based Multi-Context Guiding for Few-Shot Semantic Segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence* 33, (2019) 8441–8448. 7

[28] K. Rakelly, E. Shelhamer, T. Darrell, A. A. Efros, and S. Levine. Conditional Networks for Few-Shot Semantic Segmentation. In ICLR. 2018 . 7

[29] Z. Lu, S. He, X. Zhu, L. Zhang, Y.-Z. Song, and T. Xiang. Simpler is Better: Few-shot Semantic Segmentation with Classifier Weight Transformer. In ICCV. 2021 . 7

[30] J. Cen, P. Yun, J. Cai, M. Y. Wang, and M. Liu. Deep Metric Learning for Open World Semantic Segmentation. In 2021 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2021 . 8

[31] H. Dong, Z. Chen, M. Yuan, Y. Xie, J. Zhao, F. Yu, B. Dong, and L. Zhang. Region-Aware Metric Learning for Open World Semantic Segmentation via Meta-Channel Aggregation 2022. 8

[32] C. Zhou, F. Liu, C. Gong, T. Liu, B. Han, and W. Cheung. KRADA: Known-region-aware Domain Alignment for Open World Semantic Segmentation 2021. 9

[33] J. Myers-Dean, Y. Zhao, B. Price, S. Cohen, and D. Gurari. Generalized Few-Shot Semantic Segmentation: All You Need is Fine-Tuning 2021. 11

[34] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2016 . 12

[35] S. R. Richter, V. Vineet, S. Roth, and V. Koltun. Playing for Data: Ground Truth from Computer Games. In B. Leibe, J. Matas, N. Sebe, and M. Welling, eds., European Conference on Computer Vision (ECCV), volume 9906 of *LNCS*. Springer International Publishing, 2016 102–118. 12

[36] P. Pinggera, S. Ramos, S. Gehrig, U. Franke, C. Rother, and R. Mester. Lost and found: detecting small road hazards for self-driving vehicles. In 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). 2016 . 12

[37] H. Blum, P.-E. Sarlin, J. Nieto, R. Siegwart, and C. Cadena. The Fishyscapes Benchmark: Measuring Blind Spots in Semantic Segmentation. *arXiv preprint arXiv:1904.03215* . 12