1st Place Solution of The Robust Vision Challenge 2022
Semantic Segmentation Track

Junfei Xiao¹ Zhichao Xu³ Shiyi Lan² Zhiding Yu² Alan Yuille¹ Anima Anandkumar²
¹Johns Hopkins University ²NVIDIA ³Fudan University

Abstract

This report describes the winning solution to the Robust Vision Challenge (RVC) semantic segmentation track at ECCV 2022. Our method adopts the FAN-B-Hybrid model as the encoder and uses SegFormer as the segmentation framework. The model is trained on a composite dataset consisting of images from 9 datasets (ADE20K, Cityscapes, Mapillary Vistas, ScanNet, VIPER, WildDash 2, IDD, BDD, and COCO) with a simple dataset balancing strategy. All the original labels are projected to a 256-class unified label space, and the model is trained using a cross-entropy loss. Without significant hyperparameter tuning or any specific loss weighting, our solution ranks the first place on all the testing semantic segmentation benchmarks from multiple domains (ADE20K, Cityscapes, Mapillary Vistas, ScanNet, VIPER, and WildDash 2). The proposed method can serve as a strong baseline for the multi-domain segmentation task and benefit future works. Code will be available at https://github.com/lambert-x/RVC_Segmentation.

1. Introduction

In the past few years, advances in deep learning have led to significant progress in visual recognition. However, the robustness of state-of-the-art deep learning models remains an open issue. On the one hand, real-world applications require models to be deployed “in the wild”. On the other hand, many current deep models have been shown to be brittle to distributional shifts and natural perturbations. This phenomenon raised considerable interest in open problems such as domain generalization and adaptation.

There is rich literature in domain generalization [18, 24] where popular methods include, but are not limited to: domain randomization, domain invariant representation learning, disentanglement learning and meta learning, etc. One approach related to this work is multi-dataset training [10], in which the authors show that a simple combination of multiple datasets with label space alignment can outperform strong domain generalization approaches.

Another interesting trend is the recent surge of Vision Transformers (ViTs). Several works [1, 13, 15, 19] almost simultaneously pointed out that ViTs demonstrate surprisingly strong robustness to out-of-distribution scenarios. For example, SegFormer [19] demonstrates significantly better results over CNN-based strong methods in Cityscapes-C, a more challenging variant of Cityscapes contaminated by 16 types of natural corruption. More recently, [23] introduced the fully attentional network (FAN), a family of ViT backbones with state-of-the-art accuracy and robustness in both image classification and downstream tasks.

This report describes the winning solution to the RVC 2022 semantic segmentation track. This year, the challenge features benchmarking of a single semantic segmentation model on six datasets, spanning both indoor/outdoor and synthetic/real. Thus, it presents a great challenge to the generalization capability of a model over different domains. Our solution is inspired by the above advances in both multi-dataset training and ViTs, as will be detailed in the rest of the report.

2. Method

Backbone. We adopt FAN-B-Hybrid [23] as our backbone encoder due to its great robustness on multiple benchmarks.
Table 2. Comparison with previous methods. Measured by class mIoU. The best number in each column is highlighted in bold.

| Method Name          | Year/Rank | ADE20K | Cityscapes | Mapillary | ScanNet | VIPER | WildDash 2 |
|----------------------|-----------|--------|------------|-----------|---------|-------|------------|
| MSig1080/RVC [10]   | 2020 / 2nd| 33.18  | 80.7       | 34.19     | 48.5    | 40.7  | 34.71      |
| SN_RN152_PyrX8_RVC [2]| 2020 / 1st| 31.12  | 74.7       | 40.43     | 54.6    | 62.5  | 42.29      |
| FAN_NV_RVC (Ours)    | 2022 / 1st| **43.46**| **82.0**   | **55.27** | **58.6**| **69.8**| **47.5**   |

Table 3. Optimizer & hyper-parameters details.

| Operation             | Setting                        |
|-----------------------|--------------------------------|
| Optimizer             | AdamW [12]                     |
| Learning rate         | 6e-5                           |
| Weight decay          | 0.01                           |
| Optimizer momentum    | $\beta_1, \beta_2 = 0.9, 0.999$|
| Batch size            | 64                             |
| Learning rate schedule| Poly [4]                       |
| Warmup iters          | 1500                           |

Table 4. Training data augmentations.

| Operation             | Setting                        |
|-----------------------|--------------------------------|
| Resize                | Scale: (2048, 1024), Ratio: (0.5, 2.0) |
| RandomCrop            | Crop size: (1024, 1024)         |
| RandomFlip            | Prob: 0.5                       |
| PhotoMetricDistortion | Default                        |

Table 5. Testing data augmentations.

| Operation | Setting                        |
|-----------|--------------------------------|
| Resize    | Scale: (2048, 1024)            |
| Multi-scale| Ratios: (0.5, 0.75, 1.0, 1.25, 1.5, 1.75) |
| Flip      | True                           |

3. Implementation Details

We built our codebase with MMSegmentation [5]. The length of the training process is 80,000 iterations, while the first half training is without BDD and IDD datasets. Table 3 provides detailed information about the optimizer and hyperparameter settings. Training and testing data augmentations are detailed in Table 4 and Table 5. The model is trained on 64 V100 GPUs (32G), and the whole training procedure takes ~35 hours.

---

1 https://github.com/NVlabs/FAN
2 https://github.com/ozendelait/rvc_devkit
References

[1] Yutong Bai, Jieru Mei, Alan L Yuille, and Cihang Xie. Are transformers more robust than cnns? *Advances in Neural Information Processing Systems*, 34:26831–26843, 2021.

[2] Petra Bevandić, Marin Oršić, Ivan Grubišić, Josip Šarić, and Siniša Šegvić. Multi-domain semantic segmentation with pyramidal fusion. *arXiv preprint arXiv:2009.01636*, 2020.

[3] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018.

[4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. 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.

[5] MMISegmentation Contributors. Mmsegmentation: Openmmlab semantic segmentation toolbox and benchmark. *Available online: https://github.com/openmmlab/mmsegmentation (accessed on 18 May 2022)*, 2020.

[6] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[7] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017.

[8] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, large mini-batch sgd: Training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

[9] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2018.

[10] John Lambert, Zhuang Liu, Ozan Sener, James Hays, and Vladlen Koltun. Mseg: A composite dataset for multi-domain semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2879–2888, 2020.

[11] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[12] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.

[13] Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. *Advances in Neural Information Processing Systems*, 34:23296–23308, 2021.

[14] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulò, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *Proceedings of the IEEE international conference on computer vision*, pages 4990–4999, 2017.

[15] Sayak Paul and Pin-Yu Chen. Vision transformers are robust learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2071–2081, 2022.

[16] Stephan R Richter, Zeeshan Hayder, and Vladlen Koltun. Playing for benchmarks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2213–2222, 2017.

[17] Girish Varma, Anbumani Subramanian, Anoop Namboodiri, Mmannoh Chandraker, and CV Jawahar. Idl: A dataset for exploring problems of autonomous navigation in unconstrained environments. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1743–1751. IEEE, 2019.

[18] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[19] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jinshi Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021.

[20] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yining Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020.

[21] Oliver Zendel, Katrin Honauer, Markus Murschitz, Daniel Steininger, and Gustavo Fernandez Dominguez. Wilddash creating hazard-aware benchmarks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 402–416, 2018.

[22] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017.

[23] Daquan Zhou, Zhiding Yu, Enze Xie, Chaoxue Xiao, Animashree Anandan, Jiashi Feng, and Jose M Alvarez. Understanding the robustness in vision transformers. In *International Conference on Machine Learning*, pages 27378–27394. PMLR, 2022.

[24] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.