Improving Robustness of Semantic Segmentation Models with Style Normalization

Evani Radiya-Dixit
Department of Computer Science
Stanford University
evanir@stanford.edu

Andrew Tierno
Department of Computer Science
Stanford University
atierno@stanford.edu

Felix Wang
Department of Computer Science
Stanford University
felixw17@stanford.edu

Abstract

We introduce a novel technique for data augmentation with the goal of improving robustness of semantic segmentation models. Standard data augmentation methods rely upon augmenting the existing dataset with various transformations of the training samples but do not utilize other existing datasets. We propose a method that draws images from external datasets that are related in content but perhaps stylistically different; we perform *style normalization* on these external datasets to counter differences in style. We apply and benchmark our technique on the semantic segmentation task with the DeepLabv3+ model architecture and the Cityscapes dataset, leveraging the GTA5 dataset for our data augmentation.

1 Introduction

The task of semantic segmentation is a key topic in the field of computer vision. Recent advances in deep learning have yielded increasingly successful models [1, 3] and remarkable improvement on standard benchmark datasets Cityscapes [4] and PASCAL VOC 2012 [5]. One important goal of semantic segmentation models is robustness: the ability to successfully function on unusual or unexpected inputs. Our solution is to augment the existing dataset with images from a different source that share similar *content domains* yet perhaps vary in their *style domains*. We define the style domain of an image as the aspects of the image linked to the medium from which it originates. For example, there exists an inherent stylistic difference in photographs of the real world and those generated by a computer.

Intuitively, semantic segmentation should depend only the content of an image, and not on the style. Indeed, the style of an image captures domain-specific properties, while the content is domain-invariant. We choose to focus on the DeepLabv3+ model [3] for semantic segmentation on the Cityscapes dataset. We will apply our data augmentation technique with the GTA5 dataset [10]; we hypothesize that the addition of such synthetically generated data with style normalized to the style of the Cityscapes dataset will improve performance.

2 Related Work

Many data augmentation methods exist for semantic segmentation. One standard approach entails randomly rescaling inputs [2]. Other standard techniques include random cropping and horizontal

Preprint. Work in progress.

flipping [7], random jittering and translations [9]. We came across two other interesting approaches. The first was partitioning images into overlapping regions in [11]. The second approach was a novel technique called combinatorial cropping, where all possible combinations of ground-truth labels are used to generate pixel masks that act on all training samples. This technique was introduced in [6]. We did not come across any papers that utilized an approach similar to ours, i.e. performing data augmentation with an external dataset while normalizing style.

3 Datasets

The Cityscapes dataset collects a diverse set of street view images from 50 cities in Germany and surrounding countries. Some examples can be seen in Figure 1. Each image comes with a pixel level annotation classifying each pixel into one of 19 categories. Sample categories include person, vehicle, building, and sky. Due to computational limitations, we used $\frac{1}{25}$ of the original Cityscapes dataset.

The GTA5 dataset consists of screenshots of the open world sandbox game *Grand Theft Auto V*. The game is a particularly apt content domain as it is set in a large metropolitan city where street view style images are possible. Further, the GTA5 images have pixel level image labels that are compatible with those of Cityscapes; however, they required some additional preprocessing. The Cityscapes and GTA5 datasets have a difference in their representations of ground truth. In particular, the Cityscapes dataset encodes class labels with a grayscale image where each pixel's grayscale value represents the class label. On the other hand, the GTA5 dataset encodes class labels with an image where the pixel color represents the class label. This difference is displayed below in Figure 2.

(a) Cityscapes, image 1.

(b) GTA5, image 1.

(c) Cityscapes, image 2.

(d) GTA5, image 2.

Figure 1: Selected images from Cityscapes and GTA5 datasets.

Note the differences in the styles of the Cityscapes and GTA5 datasets. Shown above in Figure 1 are three Cityscapes images (from Aachen, Bremen, and Bochum, respectively) and three GTA5 images. There are some evident stylistic differences between the two which stem from efficiency tricks employed by GTA5's graphical engine to quickly render the screen for the player. We also note a far more vibrant palette in the GTA5 set compared to the drab appearance of the Cityscapes images. Despite these stylistic differences, both domains share a highly similar content domain: cars, trees, buildings, etc.

4 Methods

We selected the well known DeepLabv3+ architecture for semantic segmentation and used a popular PyTorch implementation (https://github.com/jfzhang95/pytorch-deeplab-xception). DeepLabv3+

(a) Cityscapes ground truth.

(b) GTA5 ground truth.

Figure 2: Ground truth examples from the Cityscapes and GTA5 datasets.

uses a pre-trained ResNet-101 model as its backbone but adds two additional modules (an atrous spacial pyramid pooling module and decoder module) designed specifically for the task of semantic segmentation. It utilizes cross entropy loss. Cross entropy loss is defined as follows: for a set of classes $\mathcal{C}$ and an image $\mathcal{I}$, if $y_{i,c}$ indicates whether the true label of pixel i is c and $\hat{y}_{i,c}$ is the probability computed by our model that pixel i is of class c then

$$CE = -\sum_{i \in \mathcal{I}} \sum_{c \in \mathcal{C}} y_{i,c} \log \hat{y}_{i,c}.$$

For style normalization we utilized a recent state-of-the-art image-to-image translation model called UNIT, introduced in [8]. We used a pretrained model designed specifically for converting images between the Cityscapes and GTA5 style domains.

(a) Input to UNIT model.

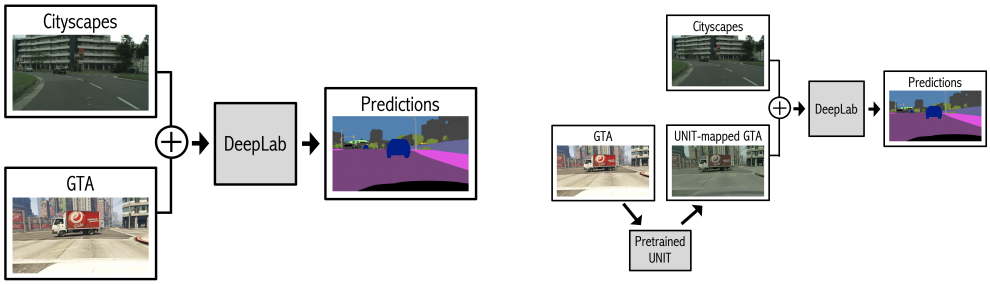
(b) Output from UNIT model.

Figure 3: UNIT translates an image from the GTA5 style domain to the Cityscapes style domain.

5 Experiments and Results

To quantify the effects of our data augmentation technique on the robustness of semantic segmentation models, we ran two primary experiments. For the first experiment we took a DeepLabv3+ network pretrained on the PASCAL VOC and Semantic Boundaries dataset, and applied transfer learning with two additional datasets: the first dataset consisted of the 587 Cityscapes training images and 302 additional GTA5 images, and the second dataset consisted of the 587 Cityscapes training images and the 302 GTA5 images mapped to the Cityscapes style domain with UNIT. For the second experiment, we trained DeepLabv3+ from scratch with two datasets: the first dataset consisted of the 587 Cityscapes training images and 587 additional GTA5 images, and the second dataset consisted of the 587 Cityscapes images and 587 additional GTA5 images mapped to the Cityscapes style domain with UNIT.

The standard benchmark statistic for semantic segmentation is mean intersection-over-union score (MIoU). Intuitively, the intersection-over-union quantifies how accurately a particular model estimates the location of an object relative to a ground truth image by computing the ratio of the number of pixels the model correctly identifies (intersection) to the total number of pixels representing either the ground truth or object or the model's prediction of the object (union). To extend this notion beyond binary classification, we introduce the notion of a confusion matrix. A confusion matrix M is defined



(a) Baseline Model

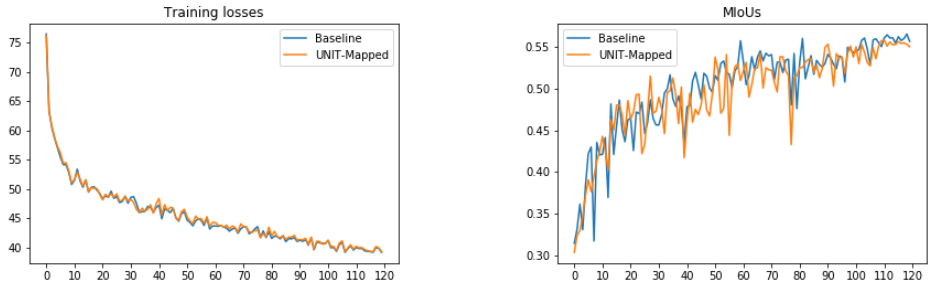
(b) UNIT-Mapped Model

Figure 4: A comparison of our data pipelines for our baseline

such that M_{ij} is the number of pixels whose ground truth label is i that the model classifies as j. Notice that the diagonal elements M_{ii} represent correctly classified pixels. Suppose we have a set of class labels $\mathcal{C}$. We can then define

$$\text{MIoU}(M) = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \frac{M_{cc}}{\sum_{i \in \mathcal{C}} M_{ci} + \sum_{i \in \mathcal{C}} M_{ic} - M_{cc}}.$$

As stated above, we first used a pretrained DeepLabv3+ model and applied transfer learning in two ways. For both models we trained on the first combined dataset of Cityscapes and GTA5, 587 images of each. For the baseline model, DeepLabv3+ was trained on this dataset to produce semantic segmentation predictions. For the UNIT-Mapped model, we first mapped the GTA images in our training dataset to the Cityscapes domain using the pretrained UNIT model. We then trained DeepLabv3+ on the Cityscapes images and these UNIT-mapped GTA images. Figure 4 shows our pipeline for the baseline model and our UNIT-Mapped Model. After training for 120 epochs, we evaluated our model on a test dataset of 98 Cityscapes images. Our training losses and MIoU scores over epochs are shown in Figure 5, and our final MIoU scores are shown in Table 1. Baseline and UNIT-Mapped performed comparably, having MIoU scores of 0.56 and 0.55, respectively.



(a) Training loss.

(b) MIoU scores during training on testing dataset.

Figure 5: Baseline and UNIT-mapped show similar training loss curves and MIoU curves for 120 epochs.

	MIoU Scores (training using transfer learning)
Baseline	0.56
UNIT-Mapped	0.55

Table 1: MIoU scores for our two models evaluated on the CityScapes dataset after 120 epochs of training using transfer learning.

We also trained DeepLab3+ from scratch on the second combined dataset of Cityscape and GTA5 images. Our baseline and UNIT-Mapped followed the pipelines in Figure 4. After training for 100

epochs, we evaluated our model on a test dataset of 98 Cityscapes images. Our final MIoU scores are shown in Table 2. UNIT-Mapped had a slightly higher MIoU score of 0.51 compare to the Baseline score of 0.48.

	MIoU Scores (training from scratch)
Baseline	0.48
UNIT-Mapped	0.51

Table 2: MIoU scores for our two models evaluated on the CityScapes dataset after 100 epochs of training from scratch.

Our code is available at the following link: https://bit.ly/2Pxnla9. The downloadable zip file includes our codebase for both the UNIT and DeepLab models.

6 Discussion

We find similar performance between baseline and UNIT-Mapped for our models trained using transfer learning. We hypothesize that the pretrained model may not be the best fit for the street city scenes of the Cityscapes and GTA datasets. The PASCAL Visual Object Classes (VOC) Dataset and the Semantic Boundaries Dataset (SBD) are different tasks than the semantic segmentation of classes in street scenes.

We also performed qualitative analyses of our results on this experiment. We compared the predicted semantic segmentations of our baseline model and the UNIT-Mapped model and find that the segmentations are similar with no noticeable differences. We hypothesize that training on a larger dataset would yield higher MIoU scores for our UNIT-Mapped model as well as more clear visual differences. Given our constrained resources and limited compute power, we were restricted to a small dataset. Our observed results in the pre-trained DeepLabv3+ experiments reinforce the fact that more data is necessary. The comparable performance on the task suggests that neither training regimen could shift the model's weights particularly far from the VOC/SBD optimum in the parameter space. The learned features for the VOC task effectively drowned out any subtleties of the street view segmentation task and the effect of our additional images. We believe that we would find more significant improvements provided more UNIT-Mapped GTA images.

Our results from our second round of experiments, where we trained DeepLabv3+ from scratch, show some potential. Here, we find that UNIT-Mapped slightly outperformed baseline. The improved MIoU scores suggest that mapping synthetic data onto the real-world domain could potentially improve the robustness of a real-world classifier.

7 Conclusion and Future Work

Our results, as discussed above, do not yield any significant conclusions regarding our novel technique for data augmentation. We note again that compute and time restrictions did not allow us to train DeepLabv3+ with sufficiently many training samples to achieve baseline results, as reported in other papers. Nonetheless, our results yield various promising avenues for future research. In particular, the superior performance of the DeepLabv3+ model with our novel technique for data augmentation (when trained from scratch) in comparison to the model with simply a combined dataset suggests that our technique could be successful if we used more training samples. Another area for future work is exploring the efficacy of our data augmentation approach across other tasks in computer vision. For instance, we would like to test our methodology on object detection and localization.

8 Contributions

There were two main tasks over the course of this project: data preprocessing and training DeepLabv3+. Evani worked on training the DeepLabv3+ model using transfer learning for our initial results. Andrew handled parts of the data preprocessing such as converting GTA5 images to

the Cityscapes style domain with the UNIT model and contributed to training DeepLabv3+ for the initial results. Felix worked on training the DeepLabv3+ codebase from scratch and some of the data preprocessing such as making the GTA5 and Cityscapes labels compatible.

References

[1] Holger Caesar, Jasper R. R. Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. *CoRR*, abs/1612.03716, 2016. URL `http://arxiv.org/abs/1612.03716`.

[2] 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, 2018.

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018.

[4] 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[5] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, jan 2015. doi: 10.1007/s11263-014-0733-5.

[6] Seunghoon Hong, Hyeonwoo Noh, and Bohyung Han. Decoupled deep neural network for semi-supervised semantic segmentation. In *Advances in neural information processing systems*, pages 1495–1503, 2015.

[7] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian D Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Cvpr*, volume 1, page 5, 2017.

[8] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. *CoRR*, abs/1703.00848, 2017. URL `http://arxiv.org/abs/1703.00848`.

[9] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[10] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *European Conference on Computer Vision (ECCV)*, volume 9906 of *LNCS*, pages 102–118. Springer International Publishing, 2016.

[11] Panqu Wang, Pengfei Chen, Ye Yuan, Ding Liu, Zehua Huang, Xiaodi Hou, and Garrison Cottrell. Understanding convolution for semantic segmentation. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1451–1460. IEEE, 2018.