Improving Robustness of Semantic Segmentation Models with Style Normalization

Evani Radiya-Dixit, Andrew Tierno, Felix Wang

{evanir, atierno, felixw17}@stanford.edu

Introduction

Motivation

One challenge to semantic segmentation models is the data having varying *style domains*. We define the style domain of an image to be aspects of the image linked to the medium from which it originates. **We examine the effects of normalizing style domains to improve the robustness of semantic segmentation models.**

Data

Cityscapes: real world images

GTA5 (Grand Theft Auto V): computer generated images

We drew 987 images of street scenes from each and partitioned them into 80/20 train-test splits. There are evident stylistic differences between the images (efficiency tricks of GTA5's graphical engine, more vibrant palette in the GTA5 images. However, the images share a content domain: cars, trees, buildings, etc.

Data Preprocessing

- Standardizing class labels (colored GTA5 ground truth images versus grayscale Cityscape ground truth images)
- Implementing transforms for GTA5 images similar to those applied to Cityscapes images (used in dataloader)

UNIT model for style normalization

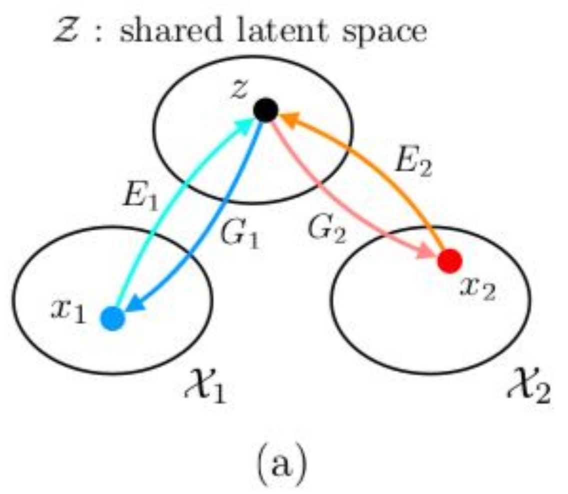

Figure 1: Shared latent latent space assumption. We assume a pair of corresponding images (x_1, x_2) in two different domains X_1 and X_2 can be mapped to a same latent code z in a shared-latent space Z. E_1 and E_2 are encoding functions, and G_1 and G_2 are generation functions.

Unsupervised Image-to-Image Translation (UNIT) converts all inputs to normalized 928 x 512 pixel images. To compare them to our larger ground truth domain images, we used cubic interpolation to upsample our UNIT mapped outputs.

Pipeline

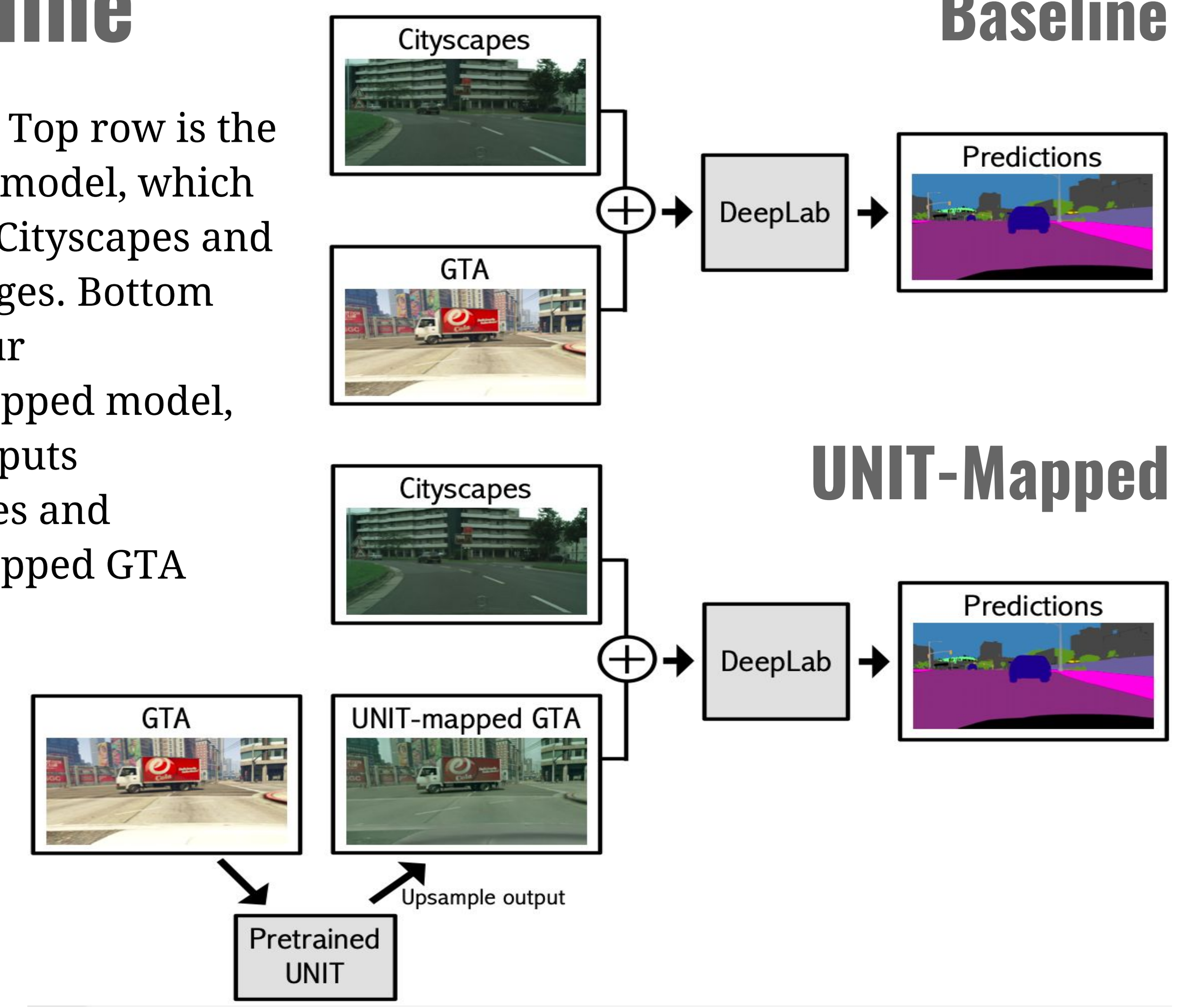

Figure 2: Top row is the baseline model, which inputs a Cityscapes and GTA images. Bottom row is our UNIT-Mapped model, which inputs Cityscapes and UNIT-Mapped GTA images.

DeepLab for semantic segmentation

DeepLabv3+ employs a re-purposed ResNet-101 for semantic segmentation by atrous convolution shown in Figure 3.

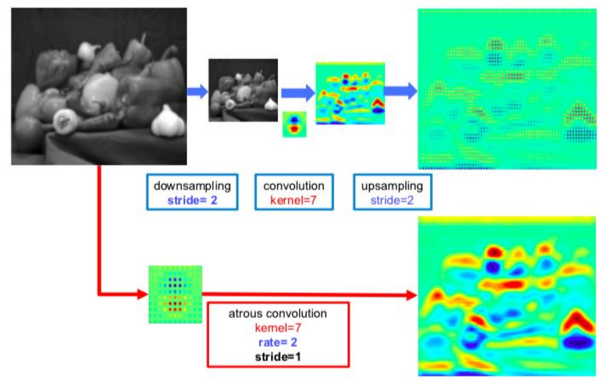

Figure 3: Top row shows sparse feature extraction with standard convolution. Bottom row shows dense feature extraction with atrous convolution.

Results

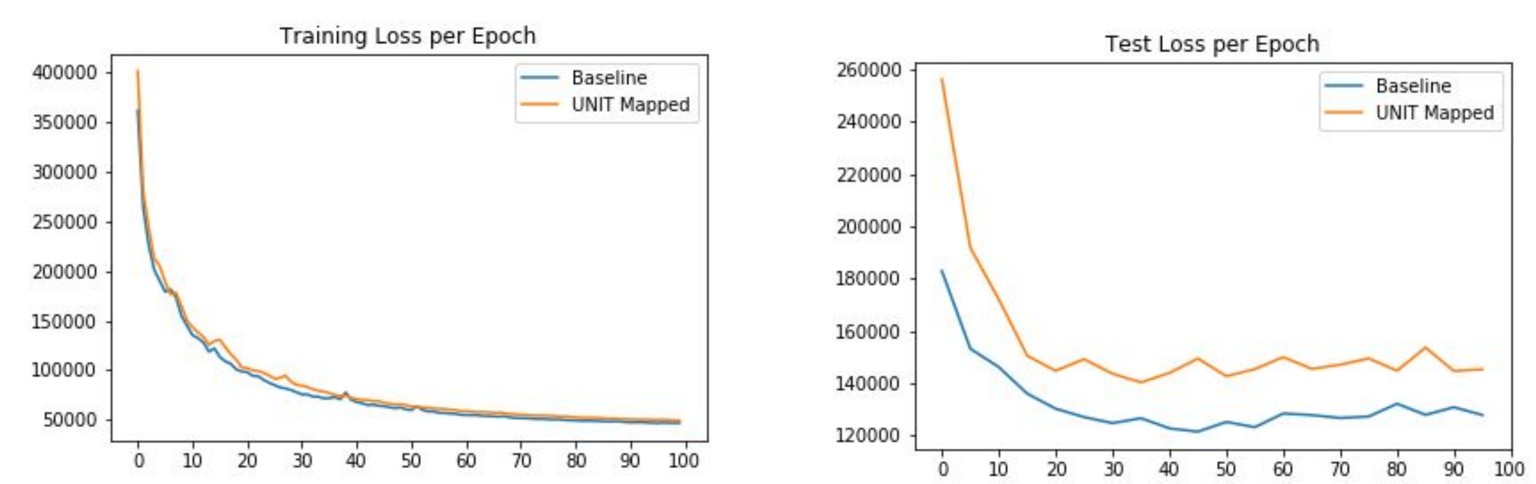

Figure 4: The training loss and testing loss per epoch evaluated on the combined dataset.

Breakdown of MIoU Scores

		Testing Dataset		
		Cityscapes	GTA5	Average
Model	Baseline	0.48	0.46	0.47
	UNIT-Mapped	0.51	0.41	0.46

Table 1: MIoU results on our baseline and experimental models evaluated on Cityscapes, GTA5, and a combination of the two.

Discussion

- UNIT-Mapped outperformed baseline on the Cityscapes semantic segmentation task, which suggests that mapping synthetic data onto the real-world domain can improve the robustness of a real-world classifier.
- UNIT-Mapped model's decreased performance on the GTA5 semantic segmentation task likely stems from accrued errors in upsampling (we visually see misalignments) and the inherently probabilistic nature of UNIT's mapping cheme.
- Style normalization does not improve performance on the combined image segmentation task

Future

- Utilizing UNIT's successor, MUNIT (Multimodal UNIT)
- Retraining UNIT to produce larger outputs, removing the need to upsample
- Testing on other synthetic databases such as Foggy Cityscapes and SYNTHIA

References

[1] 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.

[2] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. CoRR, abs/1703.00848, 2017. URL http://arxiv.org/abs/1703.00848.3