Directed Studies Research Report

Zikun Chen, 55903421
zikun.chen@alumni.ubc.ca

1 Objective

This project aims to address the domain adaptation problem for semantic segmentation. In particular, we want to develop a method that closes the domain gap between the source domain(GTA5) and target domain(CITYSCAPES) by mapping data from the source domain to the target, and train a classifier using the source data stylized as the target.

2 Datasets

Cityscapes [**?**] is a real-world dataset consisting of 3475 high quality pixel-level annotated frames from 50 different cities. I used 2975 labelled images for training, and 500 for evaluation.

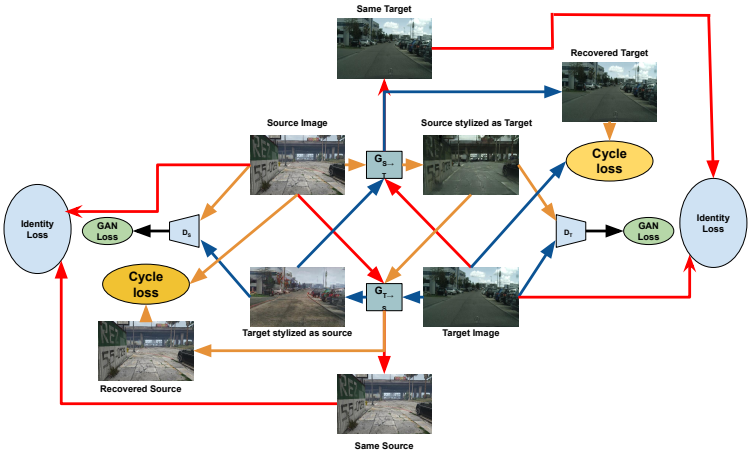
Figure 1: Sample pair of images from the CITYSCAPES data set.

GTA5 [**?**] is a synthetic dataset consisting of 24966 photo-realistic synthetic images with precise pixel-level semantic annotations. I went through the dataset and chose 16786 labelled images for training.

Figure 2: Sample pair of images from the GTA5 data set.

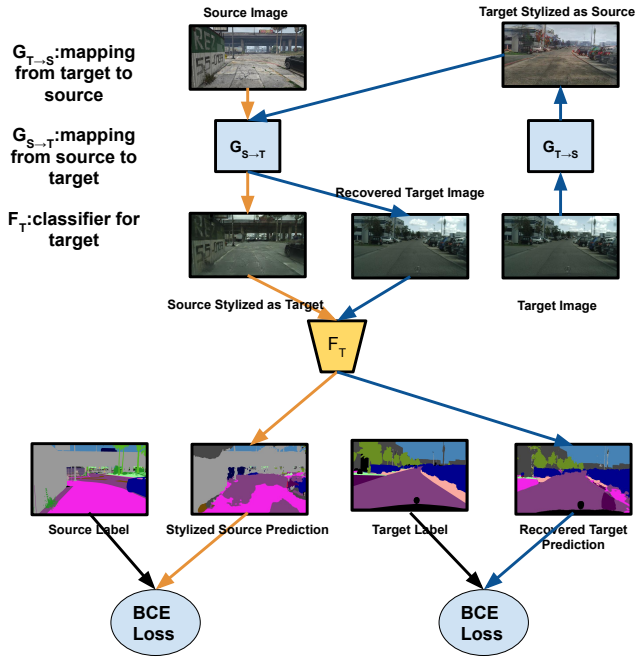
3 Models

3.1 Basic CycleGAN

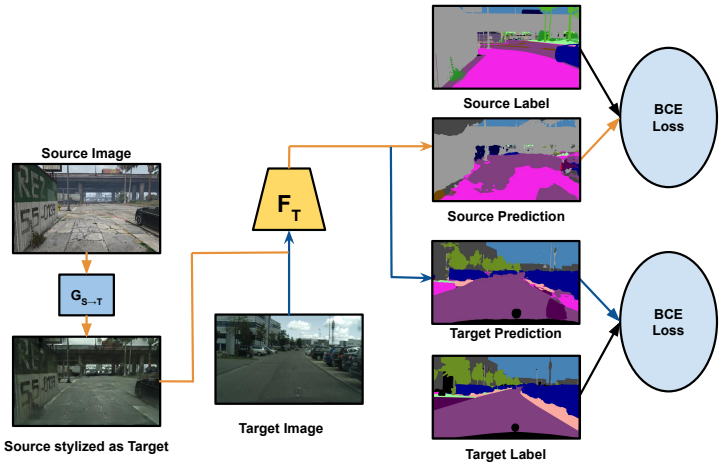


I added a classifier to preserve contents(cars, buildings) during domain transfer

3.2 Backprop from classifier to generators



3.3 Training Classifier



4 Notations

X_S : Input from the source domain (GTA5)
X_{SasT} : Input stylized as the target style (GTA5)
Y_S : Label from the source domain
X_T : Input from the target domain (CITYSCAPES)
Y_T : Input from the target domain
F_T : Classifier for the target domain
$G_{S\to T}$: Mapping from the source domain to the target
$G_{T\to S}$: Mapping from the target domain to the source
D_S : Discriminator for the source domain
D_T : Discriminator for the target domain
$CE(X, Y, F_T) := -\mathcal{L}_{(x,y)\in(X,Y)}[\sum_{k=1}^{K} \mathbb{1}_{[y=k]} \log[\sigma(F_T^k(x))]]$ cross entropy loss, X and Y are the input and the label, σ is the softmax function.

5 Loss Functions

The model learns mappings from the source domain to the target and the target to the source simultaneously. It is trained to produce samples that fools the discriminators in both domains.
$\mathcal{L}_{GAN}(G_{S\to T}, X_S, D_S, D_T) = \mathbb{E}_{x_s\in X_S}[\log D_S(x_s) + \log[1 - D_T(G_{S\to T}(x_s))]]$
$\mathcal{L}_{GAN}(G_{T\to S}, X_T, D_S, D_T) = \mathbb{E}_{x_t\in X_T}[\log D_T(x_t) + \log[1 - D_S(G_{T\to S}(x_t))]]$

The model learns to preserve contents through the cycle consistency loss and segmentation loss.
$\mathcal{L}_{Cyc}(G_{S\to T}, G_{T\to S}, X_S, X_T) = \mathbb{E}_{x_s\in X_S, x_t\in X_T}[\|G_{T\to S}(G_{S\to T}(x_s)) - x_s\|_1 + \|G_{S\to T}(G_{T\to S}(x_t)) - x_t\|_1]$
$\mathcal{L}_{G_{sem}}(G_{S\to T}, X_S, F_T, Y_S) = \mathbb{E}_{(x_s,y_s)\in(X_S,Y_S)}[CE(F_T(G_{S\to T}(x_s)), y_s)]$
$\mathcal{L}_{G_{sem}}(G_{S\to T}, G_{T\to S}, X_T, F_T, Y_T) = \mathbb{E}_{(x_t,y_t)\in(X_T,Y_T)}[CE(F_T(G_{S\to T}(G_{T\to S}(x_t))), y_t)]$

The classifier is trained on paired target data as well as paired source data stylized as target.
$\mathcal{L}_{sem}(X_{SasT}, X_T, Y_S, Y_T, F_T) = \mathbb{E}_{(x_{sast},y_s)\in(X_{SasT},Y_S),(x_t,y_t)\in(X_T,Y_T)}[CE(F_T(x_{sast}), y_s) + CE(F_T(x_t), y_t)]$

The overall loss is:

$$\mathcal{L}_{total}(X_S, Y_S, X_T, Y_T, F_T, G_{S\to T}, G_{T\to S}, D_S, D_T) = \mathcal{L}_{GAN}(G_{S\to T}, X_S, D_S, D_T) + \mathcal{L}_{GAN}(G_{T\to S}, X_T, D_S, D_T)$$
$$+\mathcal{L}_{Cyc}(G_{S\to T}, G_{T\to S}, X_S, X_T) + \mathcal{L}_{G_{sem}}(G_{S\to T}, X_S, F_T, Y_S)$$
$$+\mathcal{L}_{G_{sem}}(G_{S\to T}, G_{T\to S}, X_T, F_T, Y_T) + \mathcal{L}_{sem}(X_{SasT}, X_T, Y_S, Y_T, F_T)$$

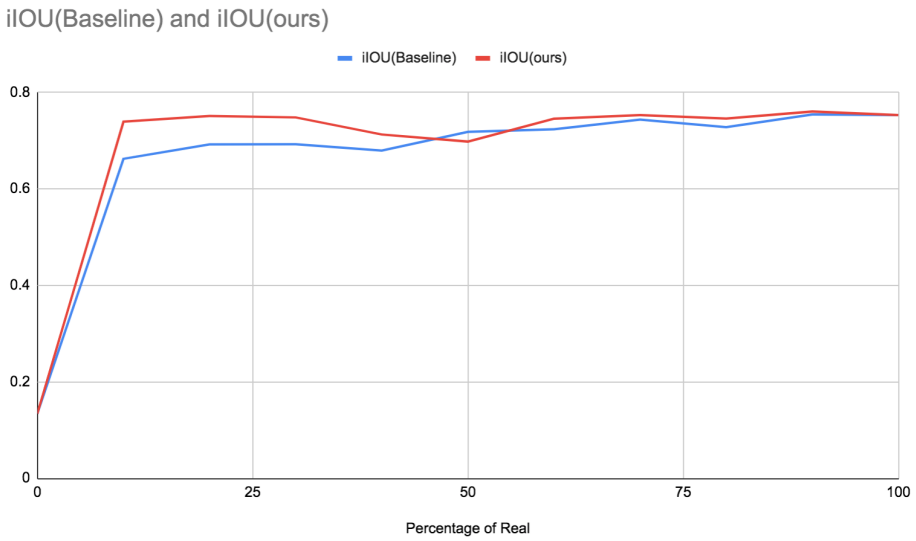
Which corresponds to solving the optimization problem

$$\underset{F_T}{\operatorname{argmin}} \min_{G_{S\to T}, G_{T\to S}} \max_{D_S, D_T} \mathcal{L}_{total}(X_S, Y_S, X_T, Y_T, F_T, G_{S\to T}, G_{T\to S}, D_S, D_T)$$

6 Experiments

6.1 Changing percentages of real(target) data used in our model

In this experiment, we kept the dataset size fixed, and only included paired data from both domains, changing the percentage each takes.
When the percentage of real data is 0%, our classifier is only trained on data from the source domain(GTA5), and tested on CITYSCAPES validation set.
When the percentage of real data is 100%, our classifier is only trained on data from the target domain(CITYSCAPES), and tested on CITYSCAPES validation set.
When the percentage is between 0 and 100%, our classifier is trained end to end with CycleGAN that stylizes data from the source domain as the target.
The baseline is a classifier with the same architecture as ours, trained without style transfer.



The evaluation metric used is weighted IOU. Our method achieved 75% accuracy with 10% of paired data from the target domain, but stops further improving as the percentage increases. Notice that the baseline reached %75 accuracy when around 90% of paired real data is used. But it's possible that the architecture for the classifier is too simple(U-Net) that it is not capable of learning more.

Issues

Some images from the GTA dataset capture scenes on the pavement instead of the road, which might lead to problems because cars don't drive on pavements in real life and the difference between pavements and roads is not significant.
Adjust the weights on loss for segmenting stylized images when increasing the percentage of source data used.
The sample is unbalanced(most instances are road and sky), adjust the weights for different class?