Conditional Generation of Satellite Images - Photorealism vs. Downstream Performance

Team: Van Anh Le, Henry Jin, Varshini Reddy, Kimon Vogt, David Harshbarger - **Partner Organization**: Microsoft AI for Good Lab

Problem Statement

High-resolution satellite images are expensive, come with privacy and licensing concerns, and large quantities are needed to train useful downstream machine learning models.

Project goal: To generate synthetic satellite images using conditional generative models.

Research question: Understanding the trade-off between photorealistic quality and downstream task performance of synthetic satellite images

Main Takeaways

1. **Synthetic images are useful as an augmentation strategy** in low-data regime, **but cannot fully replace real ones** in downstream task.
2. **No trade-off between FID and mIoU** via diversity adjustments.
3. **After post-processing step, high-diversity model output have low photorealism.**

Methodology

Data: Chesapeake Land Cover Dataset - Maryland [2]

Generative model: SPADE [1] - Pix2PixHD with VAE and without normalization of semantic mask.

Downstream model: Semantic segmentation using Unet architecture with a Resnet backbone.

Evaluation: Frechet Inception Distance (FID) for photorealism, mean Intersection-over-Union (mIoU) for downstream task performance

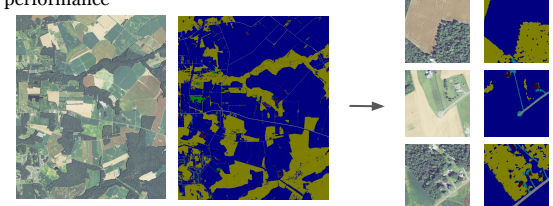


Fig. 1 Original images and labels (6,000x4,000) are split into overlapping tiles (256x256) to create training data (32k tiles) and testing data (13k tiles)

Baseline

The SPADE model trained for 4 epochs achieves **train FID of 59.09** and **test FID of 68.53**. However there is **lack of diversity** in generated output within the same class.

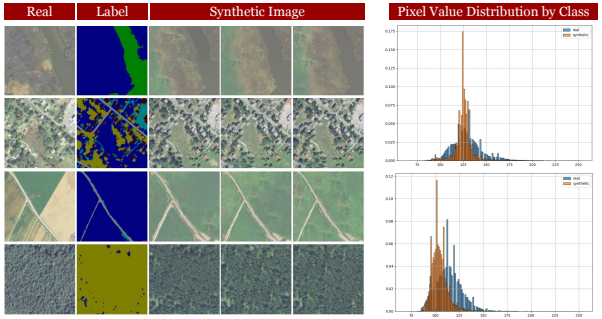


Fig. 2: Left - Baseline output of SPADE given varied input labels . Right - pixel distribution (3rd channel) of low vegetation class (blue mask) and forest class (yellow mask) showing smaller variance in synthetic images

Segmentation model trained on **100% synthetic training images** and tested on 100% of real testing images achieves **test mIoU of 0.49** compared to **0.77** for model trained on **100% real images.**

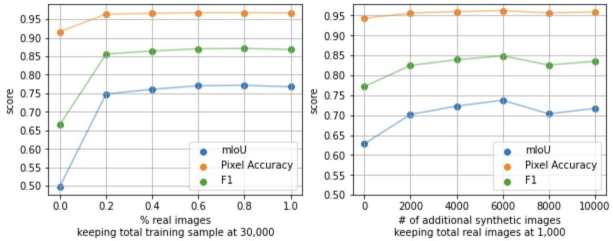


Fig. 3: Downstream performance when replacing real data with synthetic data, keeping total sample size fixe (left), and when using synthetic data to augment limited read data.

Experiment

Increase diversity of output using additional loss term in training as proposed by [3].

$$\max_G \mathcal{L}_z(G) = \mathbb{E}_{z_1,z_2}\left[\min\left(\frac{\|G(\boldsymbol{x},z_1)-G(\boldsymbol{x},z_2)\|}{\|z_1-z_2\|},\tau\right)\right]$$

FID decreases and **mIoU** increases as the as weights on diversity loss increases up to a point, suggesting that high diversity using our method reduces photorealistic quality and downstream performance of synthetic images.

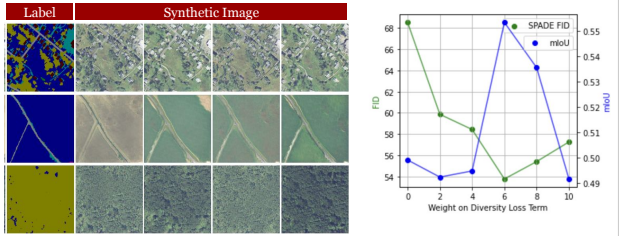


Fig. 3: Left - SPADE output with diversity loss and FID vs. mIoU trade-offs.

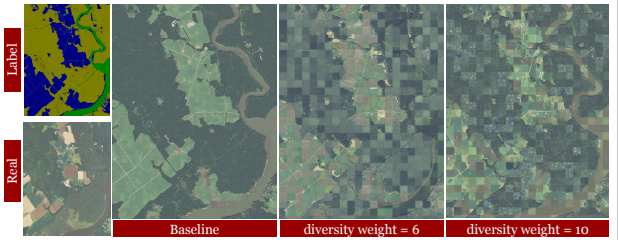


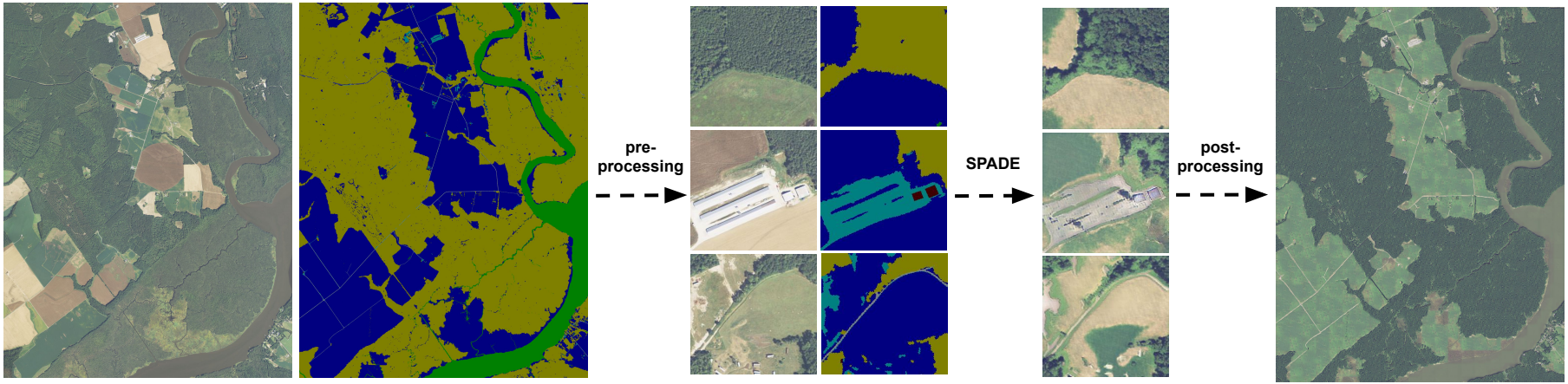
Fig. 4 Synthetic output with increasing diversity weights - Tiles are stitched back together in original order post inference

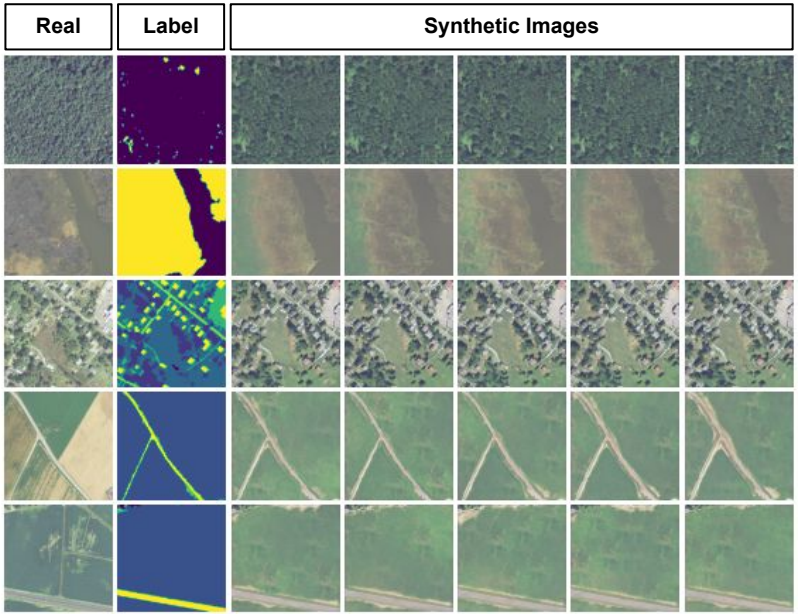
References

[1] Park, Taesung, et al. "Semantic image synthesis with spatially-adaptive normalization." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019.

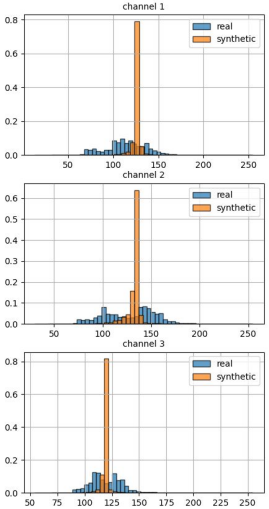
[2] Robinson C, Hou L, Malkin K, Soobitsky R, Czawlytko J, Dilkina B, Jojic N. Large Scale High-Resolution Land Cover Mapping with Multi-Resolution Data. Proceedings of the 2019 Conference on Computer Vision and Pattern Recognition (CVPR 2019)

[3] Yang, Dingdong, et al. "Diversity-sensitive conditional generative adversarial networks." *arXiv preprint arXiv:1901.09024* (2019).

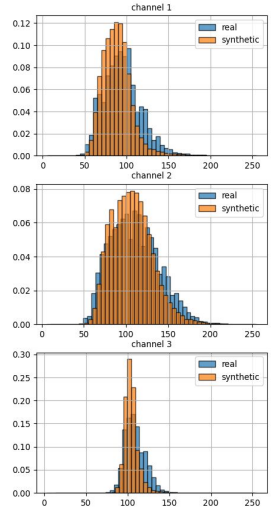

pre-
processing
SPADE
post-
processing
6000x4000
real image
6000x4000
label
256x256
training
images
256x256
training
labels
256x256
synthetic
output
6000x4000
synthetic output


Real
Label
Synthetic Images

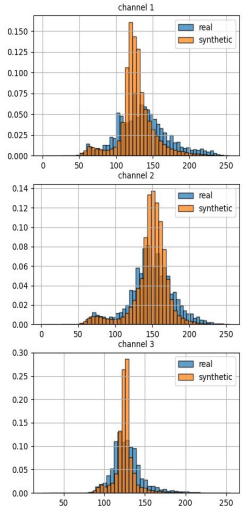
class: water - data: train



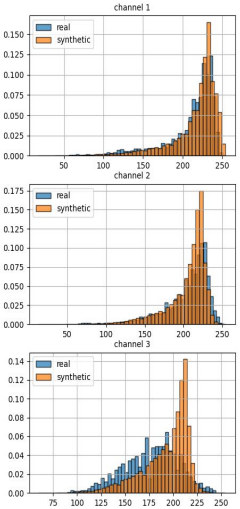
class: tree canopy/forest - data: train



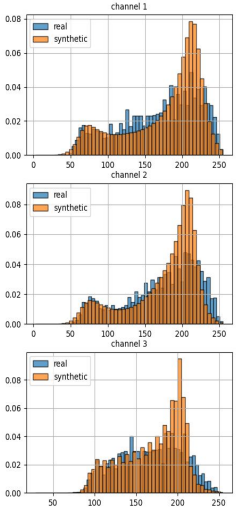
class: low vegetation - data: train



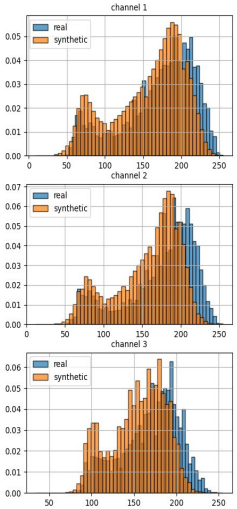
class: barren land - data: train



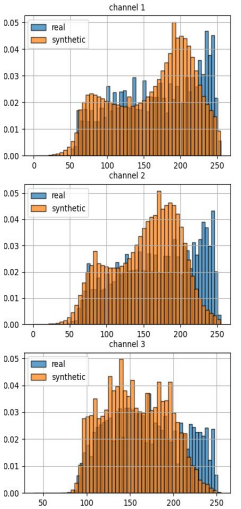
class: impervious (other) - data: train



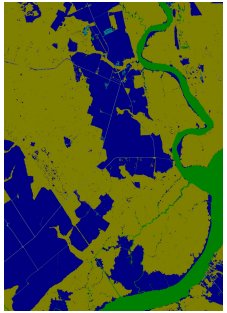
class: impervious (road) - data: train



class: building - data: train



Real Image



Label

$\lambda = 0$

$\lambda = 6$

$\lambda = 10$