# Unpaired Image Translation via Vector Symbolic Architectures

Justin Theiss[1,2], Jay Leverett[1], Daeil Kim[1], and Aayush Prakash[1]

[1] Meta Reality Labs
{theiss,jayleverett,daeilkim,aayushp}@fb.com
[2] University of California, Berkeley, CA 94720, USA

## Appendix A Additional Implementation Details

In order to encode image patches into the VSA hyperspace, we extract features using VGG19 [5] pre-trained on ImageNet [1]. We select multiple layers from VGG19 and concatenate features that share the same receptive field. For all experiments, we extract features from multiple convolutional layers in VGG19. For GTA to Cityscapes, we concatenate features for patch sizes $16 \times 16$, $8 \times 8$, $4 \times 4$, $2 \times 2$, and, $1 \times 1$. This means that extracted features within each patch are concatenated into a single feature vector, which is reduced in dimensionality using locality sensitive hashing as described in Section 3.3 of the main paper. For all experiments, we used a hypervector dimensionality of 4096, which is the dimensionality of the input to the discriminator and mapping networks. We then used 1024 channels in all layers except the last (1 channel) for the discriminator, and 4096 channels for both layers in the mapping network.

For Google Map to Aerial Photo, we use the same patch sizes as for GTA to Cityscapes. However, for Aerial Photo to Google Map, we empirically found that using "dilated" patches gave better performance. For this experiment, we use features extracted from four convolutional layers with patch sizes $32 \times 32$, $16 \times 16$, $8 \times 8$, and $4 \times 4$, where each patch has spaces between locations in the extracted feature maps (i.e., equivalent to a convolutional dilation factor of 3).

For all experiments, we trained using $256 \times 256$ images (as was done in the baseline methods). In the case of GTA to Cityscapes, we first resize the images to a height and width of $256 \times 512$ followed by random cropping to $256 \times 256$. For the Google Map and Aerial Photo datasets, images are resized from $600 \times 600$ to $256 \times 256$. Note that we do not report results for EPE in Table 1 of the main paper as we are not able to replicate their method using $256 \times 256$ image resolution for GTA to Cityscapes and it is not feasible to use their method for datasets without G-buffers. As was done in [3], we train for 20 epochs for GTA to Cityscapes and 400 epochs for Google Map to Aerial Photo and Aerial Photo to Google Map experiments. We multiply the learning rates (defined in Section 4.1 of the main paper) by a factor of 0.5 every 100K iterations.

## Appendix B Difference between VSA-Based Cyclic Loss and Perceptual Loss

Many image-to-image translation methods have included a so-called "perceptual loss" to minimize the distance between translated and source images in VGG feature space. Although our VSA-based cyclic loss (Eqn 6) may seem similar to the perceptual loss, our method is distinct: we minimize the distance between source hypervectors and translated hypervectors that are mapped back from the target to source domain, whereas methods using a perceptual loss directly minimize the distance between source and translated features. Furthermore, as noted previously, encoding these source and translated features into the VSA hyperspace gives greater assurance that different semantic content will be almost orthogonal to each other.

## Appendix C Additional Ablation Studies

As indicated above, we changed the patch sizes used during feature extraction across different experiments (i.e., "dilated" patches in Aerial Photo to Google Map). We therefore, also looked at the effect of patch size within a specific experiment. In addition to the $16 \times 16$ patch size used in the GTA to Cityscapes experiment, we also tested $32 \times 32$ and $8 \times 8$ patch sizes. Compared to the results in Table 1 (mIoU 30.89), we observed mIoU of 32.29 and 31.47 for 32 and 8 patch sizes, respectively.

Furthermore, we conducted an ablation of the weight ($\lambda$) for our VSA-based cyclic loss, showing that $\lambda < 5$ results in semantic flipping, which is reflected in lower mIoU for the GTA to Cityscapes experiment. Specifically, we observed lower mIoU for smaller (21.55 and 26.49 for $\lambda = 1, 2$) compared to larger $\lambda$ (29.13, 30.89, 29.78 for $\lambda = 5, 10, 20$).

Finally, we also tested the effect of the hypervector dimensionality. As demonstrated in [4], the hypervector dimensionality should be at least greater than 700 (the probability of two randomly sampled vectors being orthogonal approaches 1). We tested 512 and 2048 dimensional hypervectors, demonstrating worse performance for 512 compared to 2048 (29.35 vs. 30.87 mIoU for GTA to Cityscapes, respectively).

## Appendix D Hypervector Adversarial Loss

Since our hypervector adversarial loss (Eqn 5) has two negative targets and one positive target, it's possible that this could lead to an imbalance in training. For the current experiments, we have tested using a balanced weighting of the loss terms and did not empirically observe differences in stability or performance. However, this may be an important consideration for future research with other datasets.

## Appendix E Quantitative Comparison to Additional Methods for GTA to Cityscapes

In order to compare our method against other relevant methods that have also been applied to GTA to Cityscapes, we trained semantic segmentation networks using our translated GTA images as done in other methods. In these comparisons we used ResNet-50 as the feature extractor rather than VGG19. Specifically, we compared our method against CyCADA [2] and Fourier Domain Adaptation (FDA) [7]. CyCADA [2] is a method that uses the cycle-consistency approach from CycleGAN [8] in order to ensure semantic consistency between segmentation predictions before and and after image translation. Unlike GAN-based methods, FDA [7] computes the Fourier Transform of source and target images, and replaces the amplitudes of low-frequency features in the source images with those from the target images. We observed comparable performance to these approaches (45.7 mIoU for VSAIT vs. 42.7 in CyCADA [2] and 44.6 in FDA [7]).

We further compared our results to those from [6], which is a recent video-to-video translation technique that aims to prevent semantic flipping via pseudo-supervision with optical flow from sequential video frames. By using a checkpoint from the authors to evaluate semantic segmentation performance (GTA to Cityscapes, Table 1), we observed lower performance compared to our method (pixel accuracy: 66.04, class accuracy: 38.42, and mIoU: 24.53).

## Appendix F VSA Hypervector Encoding

As noted in the main paper, VGG19 works well for most natural image domains; however, it may not be optimal for encoding semantic segmentation masks or similar types of images (e.g., Google Map images). For example, using VSAIT to translate semantic masks to Cityscapes images using the same sub-sampling method used in [3], we observed worse performance compared to that reported in [3] (pixel accuracy: 69.28, class accuracy: 31.37, mIoU: 21.59). Therefore, an important future direction for this research is to understand and improve encoding methods when using VSA for computer vision applications.

## Appendix G Additional Visual Examples

### G.1 GTA → Cityscapes

In addition to the comparisons shown in the main paper, we provide more qualitative examples showing our method relative to the baseline methods for GTA to Cityscapes. As described in the main paper, GTA and Cityscapes have semantics that are naturally different. Most notably, Cityscapes has more vegetation in the upper half of the image whereas GTA has more sky (see also Figure 1 from the main paper). Therefore, many GAN-based image translation methods hallucinate trees in the sky as shown in Figures 1 and 2. More interestingly, EPE does not hallucinate trees but instead removes palm trees from images. This is

likely due to the fact that palm trees do not appear in Cityscapes. On the other hand, our method correctly retains the palm trees without hallucinating more trees in the sky.

Figure 3 provides a similar example of the typical hallucinations that occur when translating GTA images with broad sky regions. However, notice the sandy hill on the left side of the image, which we highlight in the figure. As shown in the top-right panel, the semantic segmentation label for this region is the same for both the grassy and sandy portions of the hill. Since EPE uses these labels as part of its task-level consistency loss, it generates the same feature (i.e., grass) across the region, irrespective of the underlying specific content differences.

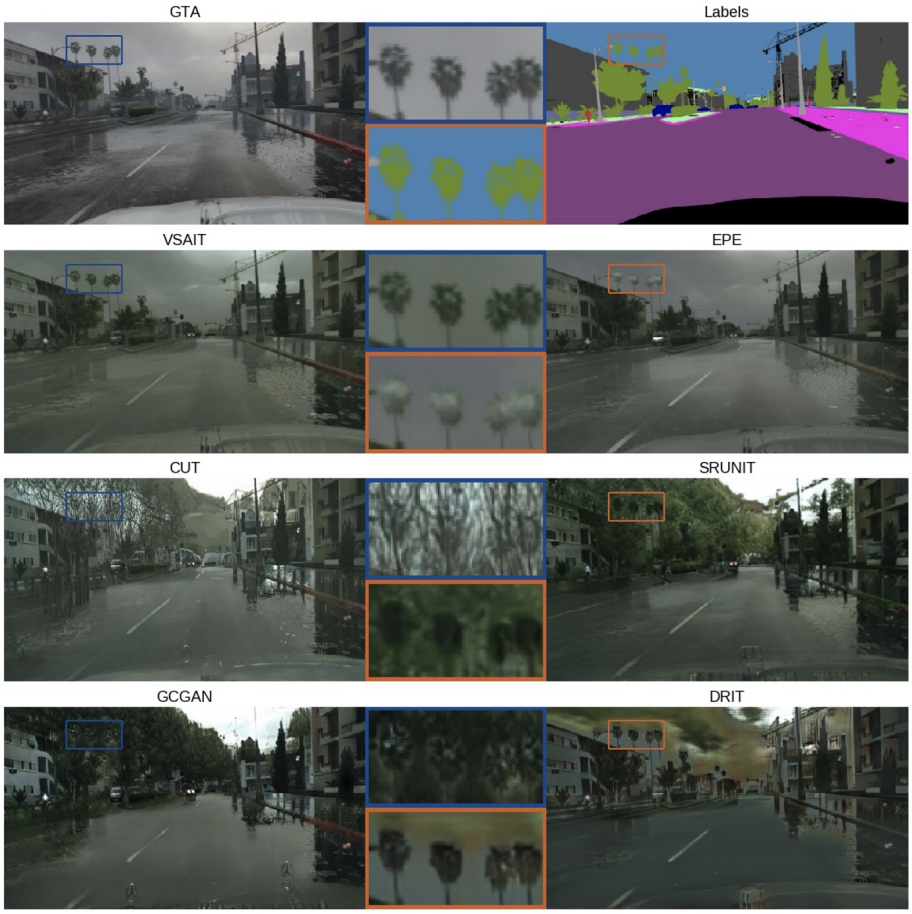


Fig. 1: Examples of semantic flipping for baseline methods in GTA to Cityscapes experiment.

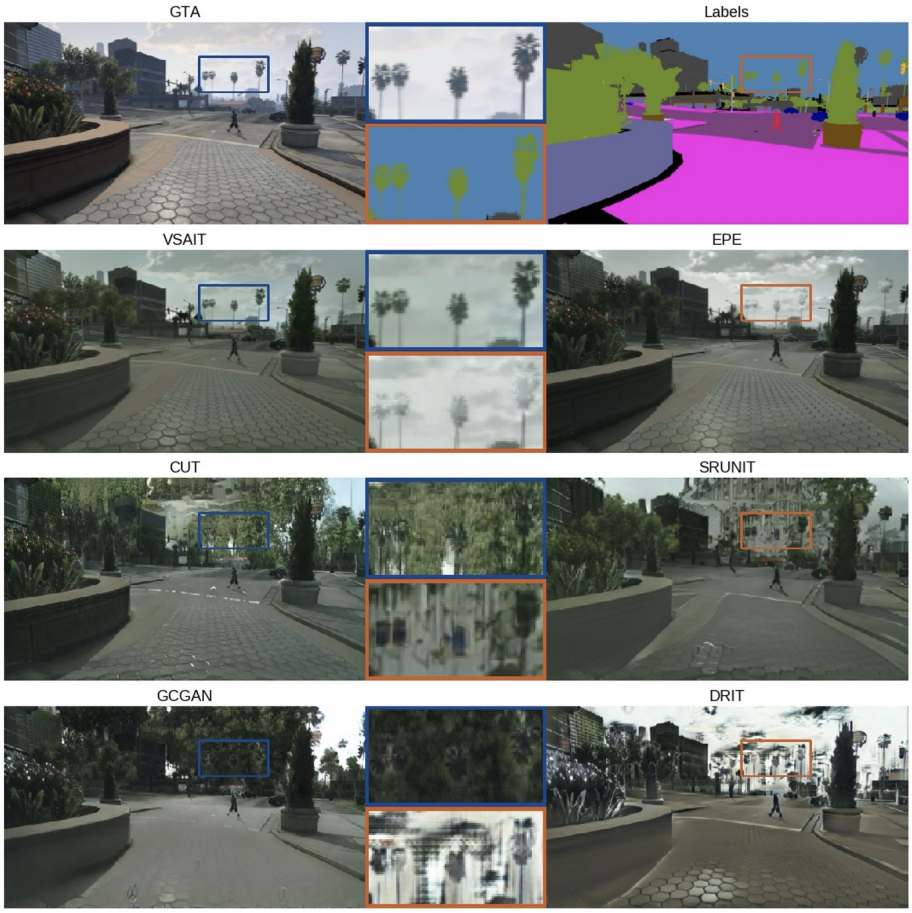


Fig. 2: Examples of semantic flipping for baseline methods in GTA to Cityscapes experiment.

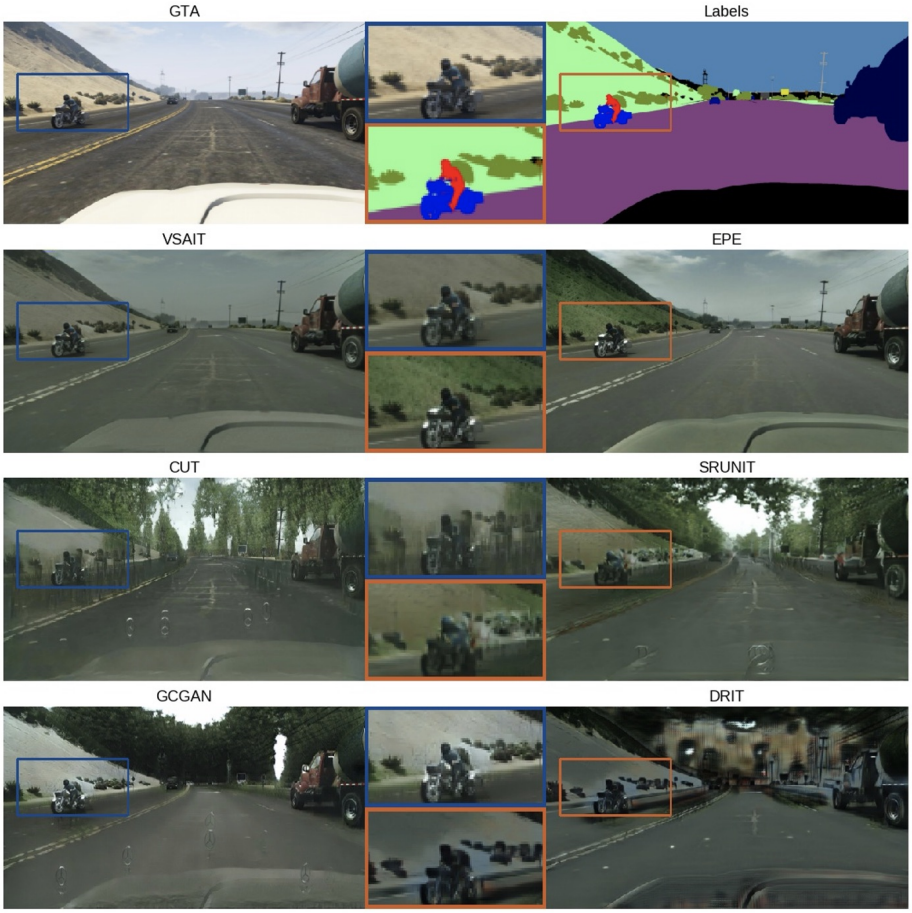


Fig. 3: Examples of semantic flipping for baseline methods in GTA to Cityscapes experiment.

### G.2 Google Map → Aerial Photo

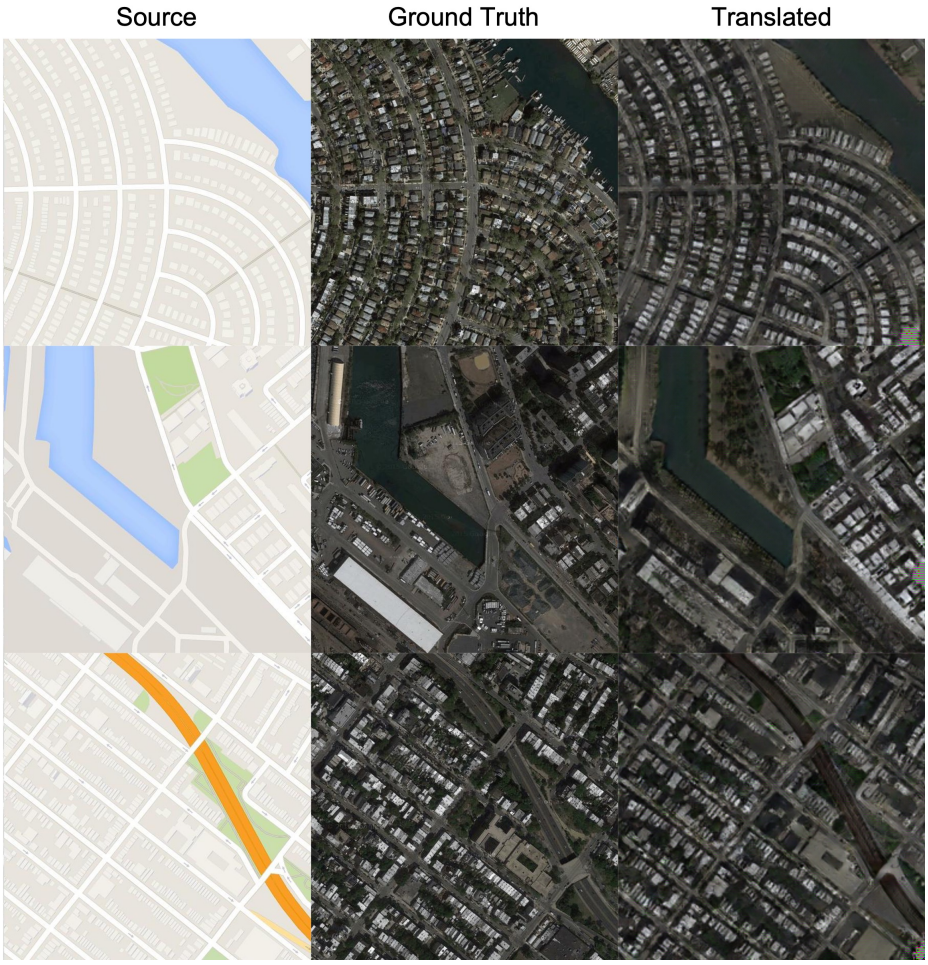


Fig. 4: Example image translations using VSAIT for the Google Map to Aerial Photo experiment.

As described in Section 4.5 of the main paper, we sub-sample the Google Map and Aerial Photo datasets (following [3]) to obtain unpaired source and target datasets that differ in their semantic statistics. As a result, the sub-sampled datasets differ in the proportion of image area containing land versus water. Therefore, similar to the GTA to Cityscapes experiment, we would expect hallucinations of houses or trees in the water. We demonstrate across multiple examples in Figures 4 and 5 that our method does not exhibit semantic flipping. It's

possible that some methods may rely on a one-to-one mapping using the RGB values in the map images (e.g., blue for water) that could reduce the diversity of translations overall. However, as shown in Figure 6, we observe diverse translations for the same RGB values across different map images. Specifically, notice the translations associated with gray pixels in each source image as highlighted in the figure.


Fig. 5: Example image translations using VSAIT for the Google Map to Aerial Photo experiment.


Fig. 6: Example image translations using VSAIT for the Google Map to Aerial Photo experiment.

### G.3 Aerial Photo → Google Map

Finally, we demonstrate examples of our image translations from the Aerial Photo to Google Map experiment. As mentioned previously, we use the same sub-sampling procedure to obtain datasets with unmatched semantic statistics. Since the sub-sampled Google Map dataset has more regions with water relative to the sub-sampled Aerial Photo dataset, we would expect that GAN-based methods might suffer from semantic flipping by hallucinating water. However, as shown in Figures 7 and 8, our method learns the correct mapping between water and does not exhibit semantic flipping in scenes without water (Figure 9).

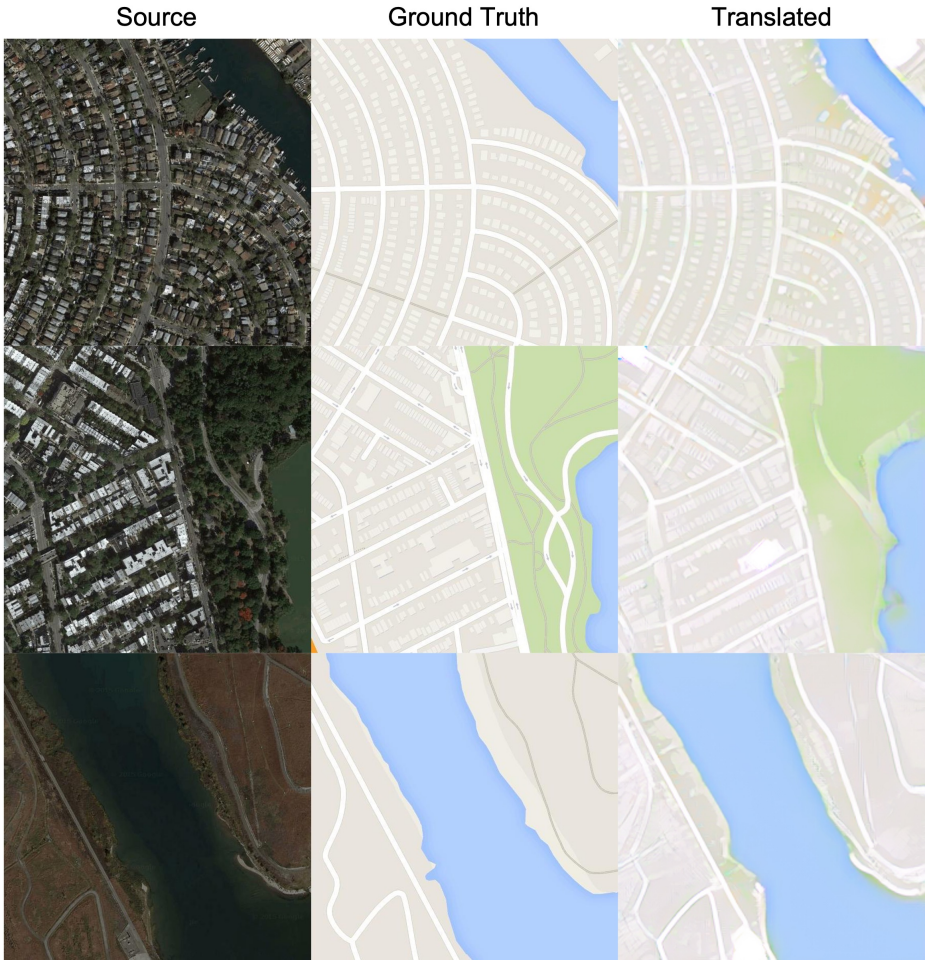


Fig. 7: Example image translations using VSAIT for the Aerial Photo to Google Map experiment.

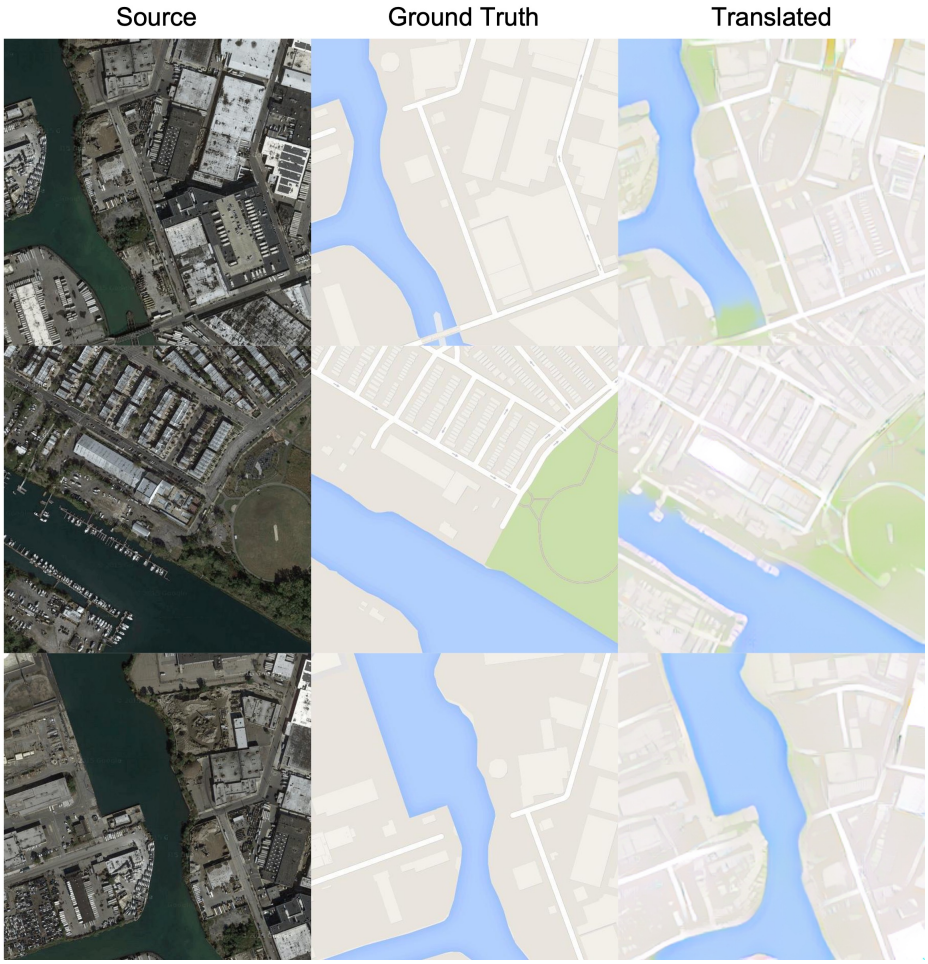


Fig. 8: Example image translations using VSAIT for the Aerial Photo to Google Map experiment.

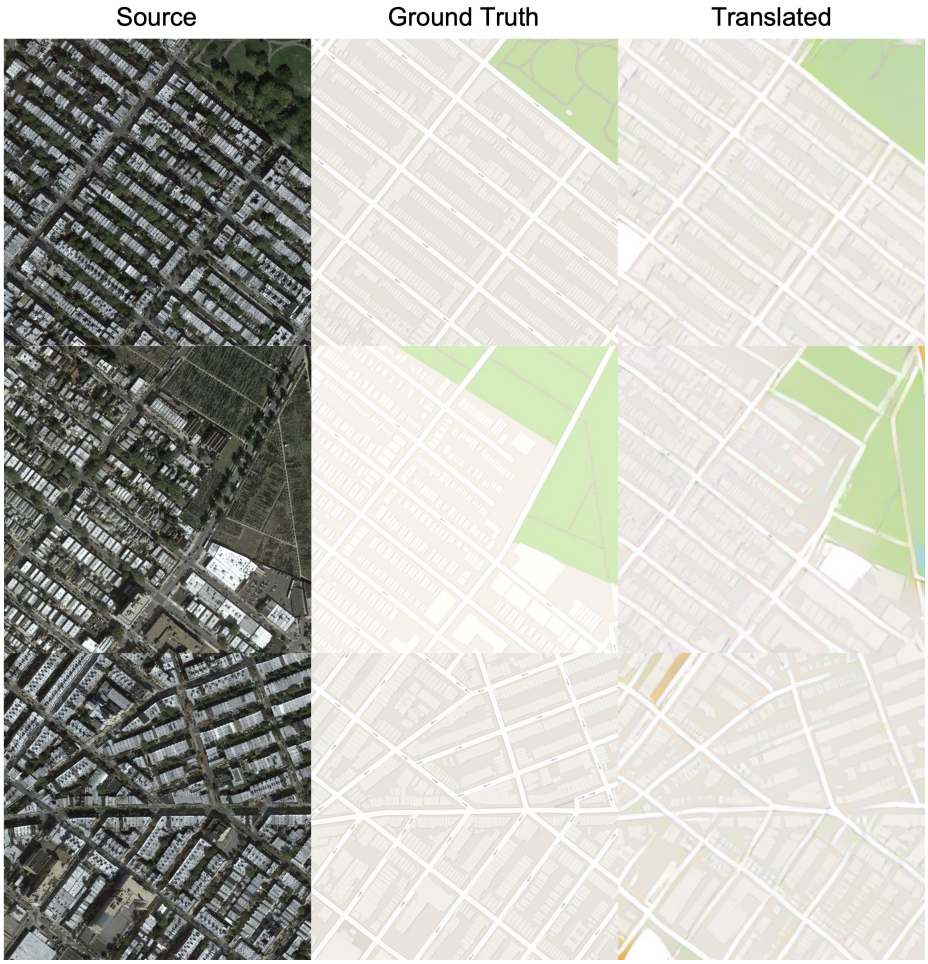


Fig. 9: Example image translations using VSAIT for the Aerial Photo to Google Map experiment.

### G.4 Visual Comparison to Baseline Methods

We additionally demonstrate the comparison between VSAIT and baseline methods for the Aerial Photo to Google Map and Google Map to Aerial Photo experiments in Figure 10. As shown in the figure, for these more difficult experiments where the datasets are specifically sub-sampled to reduce content overlap between domains, most of the baseline methods exhibit considerable semantic flipping (e.g., water to land).

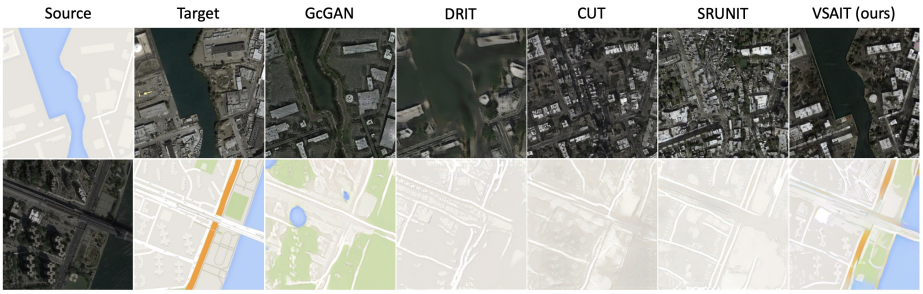


Fig. 10: Example image translations for VSAIT and baseline methods for Aerial Photo to Google Map and Google Map to Aerial Photo experiments. Note that EPE cannot be evaluated for these experiments.

## References


1. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
2. Hoffman, J., Tzeng, E., Park, T., Zhu, J.Y., Isola, P., Saenko, K., Efros, A., Darrell, T.: Cycada: Cycle-consistent adversarial domain adaptation. In: International conference on machine learning. pp. 1989–1998. PMLR (2018)
3. Jia, Z., Yuan, B., Wang, K., Wu, H., Clifford, D., Yuan, Z., Su, H.: Semantically robust unpaired image translation for data with unmatched semantics statistics. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14273–14283 (2021)
4. Neubert, P., Schubert, S., Protzel, P.: An introduction to hyperdimensional computing for robotics. KI-Künstliche Intelligenz **33**(4), 319–330 (2019)
5. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
6. Wang, K., Akash, K., Misu, T.: Learning temporally and semantically consistent unpaired video-to-video translation through pseudo-supervision from synthetic optical flow. arXiv (2022)
7. Yang, Y., Soatto, S.: Fda: Fourier domain adaptation for semantic segmentation. In: CVPR (2020)
8. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)