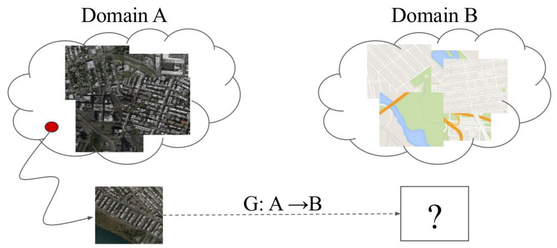
Adversarial Self-Defense for Cycle-Consistent GANs

Dina Bashkirova[1], Ben Usman[1] and Kate Saenko[1 2]
[1]Boston University [2]MIT-IBM Watson AI Lab

BOSTON UNIVERSITY

Unsupervised Image-to-Image Translation



Problem: Self-adversarial Attack

Figure 1. When incorporated into the MUNIT model, the cycle consistency property results in much more accurate translation, but it also forces the model to perform a self-adversarial attack

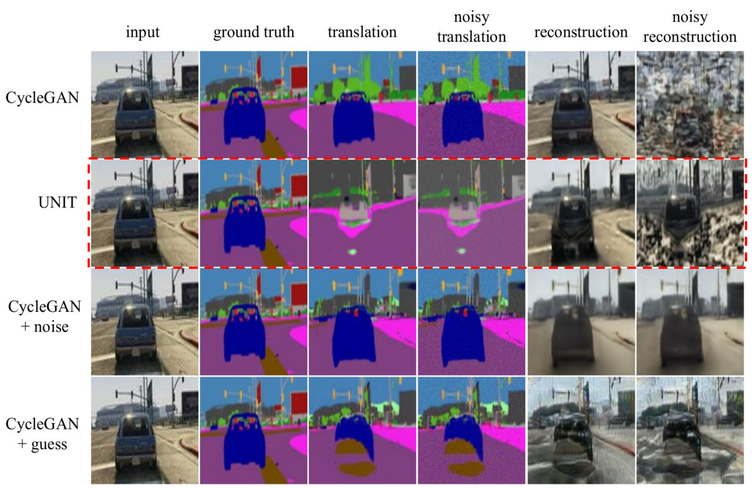


Figure 2. CycleGAN and UNIT are able to perfectly reconstruct the input image despite the semantic inaccuracy of the translations. Adding low-amplitude Gaussian noise ($\mu = 0$, $\sigma = 0.08$) to the translated image completely destroys the reconstruction. Our defense techniques (rows 3 and 4) rely on the translated image instead of embedded noise and are more robust to random perturbations.

Our defense techniques

Original cycle consistency loss

$$\mathcal{L}_{rec} = \|F(G(x)) - x\|_1$$

Adversarial training with noise

The intuition behind this approach is that adding random noise of amplitude similar to the hidden signal disturbs the embedded message. This results in high reconstruction error, so the generator cannot rely on the embedding.
The modified noisy cycle-consistency loss can be described as follows:

$$\mathcal{L}^{noisy}_{rec} = \|F(G(x) + \Delta(\theta_n)) - x\|_1 ,$$

Guess discriminator

We treat the reconstructions as the adversarial examples of the corresponding input samples, we randomly shuffle and concatenate them and then pass them to the additional *guess discriminator* that aims to correctly predict which of the images is real.
The guess discriminator loss or *guess loss* can be described as:

$$\mathcal{L}_{guess} = \begin{cases} G^A_{guess}\{\mathbf{X}, F(G(\mathbf{X})\}, & \text{with probability 0.5} \\ 1 - G^A_{guess}\{F(G(\mathbf{X})),\ \mathbf{X}\}, & \text{with probability 0.5} \end{cases}$$

where $X \sim P_A$, $G^A_{guess}(\mathbf{X}, \hat{\mathbf{X}}) \in [0, 1]$.

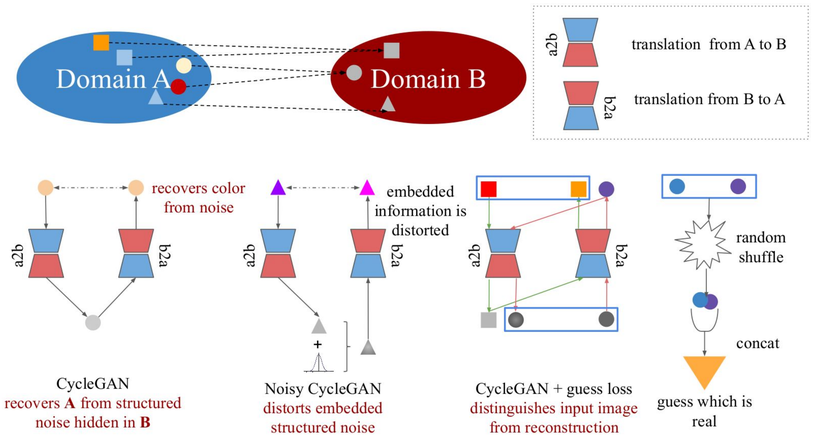


Figure 3. Comparison of architectures of the original CycleGAN and CycleGAN with our defense methods: additive noise and guess discriminator.

Our metrics

Translation quality

For any ground truth pair A_i, B_i, the one-to-many translation quality is computed as $IoU(\text{pix}(G_A(B_i)), \text{pix}(A_i))$, where pix(*) is the translation with Pix2pix from A to B. The "honest reconstruction" is compared with the supervised Pix2pix translation of the ground truth image A_i instead of the ground truth image itself in order to take into account the error produced by the Pix2pix translation.

Reconstruction honesty

$$RH = \frac{1}{N}\sum_{i=1}^{N}\{\|G_A(\lfloor G_B(X_i)\rfloor) - Y_i\|_2 - \|G_A(G_B(X_i)) - Y_i\|_2\},$$

where $\lfloor * \rfloor$ is a quantization operation, G_B is a many-to-one mapping, (X_i, Y_i) is a ground truth pair of examples from domains A and B.

Sensitivity to noise

$$SN(\sigma) = \frac{1}{N}\sum_{i=1}^{N}\|G_A(G_B(X_i) + \mathcal{N}(0, \sigma)) - G_A(G_B(X_i))\|_2$$

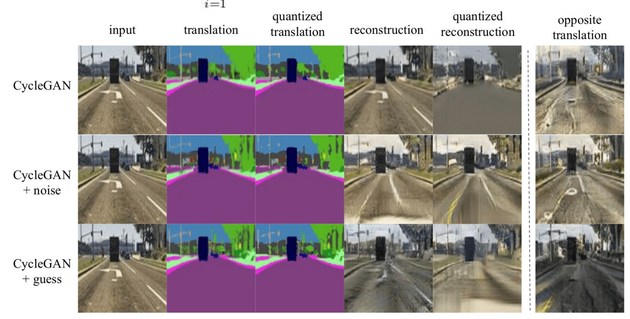


Figure 5. Quantized reconstruction results of the original CycleGAN, CycleGAN with noise defense and CycleGAN with guess loss defense. The last column represents the translation from the corresponding ground truth semantic segmentation map to real frame for comparison.

Results

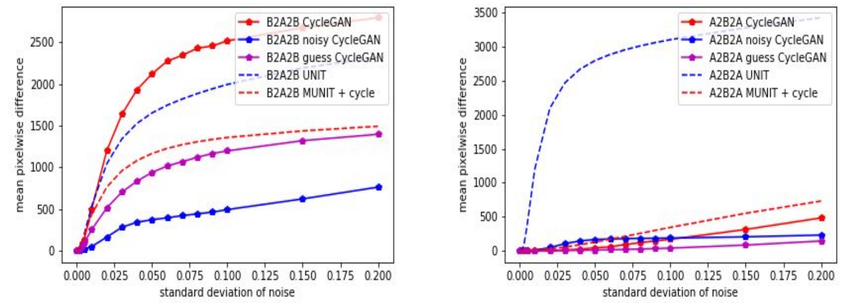


Figure 4. Illustration of sensitivity of cycle-consistent translation methods to high-frequency perturbations in one-to-many (left) and in many-to-one (right) cases. Here the domains A and B are segmentation maps and GTA video frames respectively.

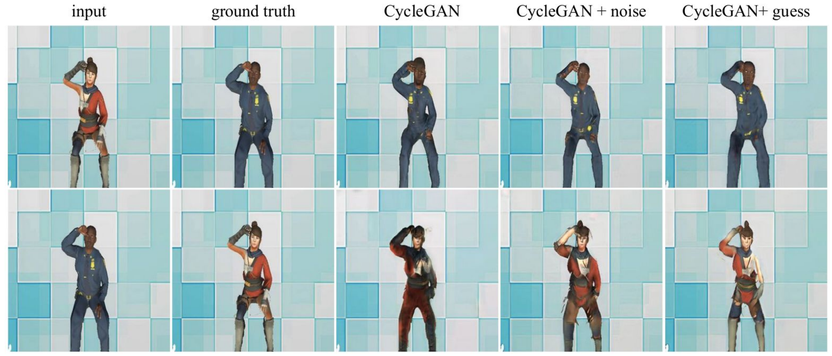


Figure 6. One-to-one translation results on SynAction dataset.

Method	acc. segm ↑	IoU segm↑	IoU p2p↑	RH↓	SN↓
CycleGAN	0.230	0.16	0.20	27.43 ± 6.14	446.92
CycleGAN + noise*	**0.240**	0.17	0.23	9.17 ± 7.37	**94.15**
CycleGAN + guess*	0.237	**0.17**	0.21	11.38 ± 7.03	212.59
CycleGAN + guess + noise*	0.236	**0.17**	**0.24**	$\mathbf{6.1} \pm \mathbf{5.85}$	150.55
UNIT	0.080	0.04	0.06	6.37 ± 11.69	361.52
MUNIT + cycle	0.131	0.08	0.17	2.5 ± 8.86	244.95
pix2pix (supervised)	0.4	0.34	–	–	–

Table 2. Results on the GTA V dataset. *acc. segm* and *IoU segm* represent mean class-wise segmentation accuracy and IoU, *IoU p2p* is the mean IoU of the pix2pix segmentation of the segmentation-to-frame mapping; *RH* and *SN* are the quantized reconstruction honesty and sensitivity to noise of the many-to-one mapping (B2A2B) respectively. * -- our proposed defense methods. The reconstruction error distributions plots can be found in the supplementary material (Section 2).

Method	acc. segm↑	IoU segm↑	IoU p2p↑	RH ↓	SN↓
CycleGAN	0.233	0.175	0.210	21.77 ± 5.16	251.19
CycleGAN + noise*	0.242	0.187	0.218	12.27 ± 4.42	**222.18**
CycleGAN + guess*	0.241	0.184	**0.224**	7.47 ± 2.38	235.43
CycleGAN + guess + noise*	**0.249**	**0.191**	0.218	$\mathbf{-0.45} \pm \mathbf{2.26}$	238.25
UNIT	0.212	0.153	0.124	19.63 ± 6.07	528.22
MUNIT + cycle	0.153	0.094	0.124	21.43 ± 7.85	687.27
pix2pix (supervised)	0.301	0.234	–	–	–

Table 3. Results on the Google Maps dataset. The notation is same as in the Table 1.

Acknowledgements

This project was supported in part by NSF and DARPA.

References

- Richter, Stephan R., et al. "Playing for data: Ground truth from computer games." *European conference on computer vision*. Springer, Cham, 2016.
- Zhu, Jun-Yan, et al. "Unpaired image-to-image translation using cycle-consistent adversarial networks." *Proceedings of the IEEE international conference on computer vision*. 2017.
- Liu, Ming-Yu, Thomas Breuel, and Jan Kautz. "Unsupervised image-to-image translation networks." *Advances in neural information processing systems*. 2017.
- Huang, Xun, et al. "Multimodal unsupervised image-to-image translation." *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018.
- Chu, Casey, Andrey Zhmoginov, and Mark Sandler. "CycleGAN, a master of steganography." *arXiv preprint arXiv:1712.02950* (2017).
- Sun, Ximeng, Huijuan Xu, and Kate Saenko. "A Two-Stream Variational Adversarial Network for Video Generation." *arXiv preprint arXiv:1812.01037* (2018).