Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks

(CycleGAN)

Jun-Yan Zhu* Taesung Park* Phillip Isola Alexei A. Efros
Berkeley AI Research (BAIR) laboratory, UC Berkeley

ICCV 2017

Slides compiled by Lars Gjesteby

August 9, 2018

Motivation

- Image-to-image translation between domains (i.e. art style transfer, season transfer, animal transfiguration)
- Paired training data not needed
- Learn domain-level relationships

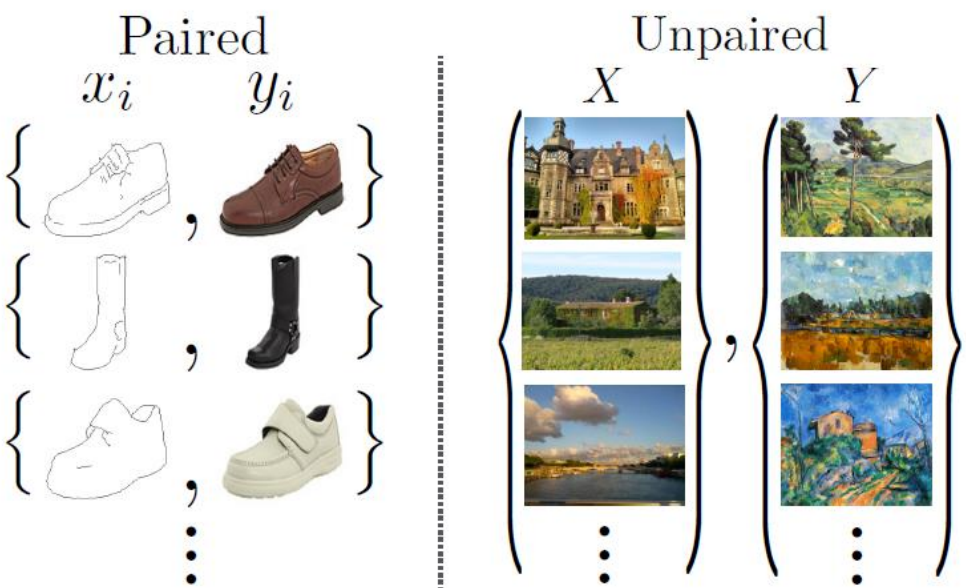


Figure 2: *Paired* training data (left) consists of training examples $\{x_i, y_i\}_{i=1}^N$, where the correspondence between x_i and y_i exists [21]. We instead consider *unpaired* training data (right), consisting of a source set $\{x_i\}_{i=1}^N$ ($x_i \in X$) and a target set $\{y_j\}_{j=1}$ ($y_j \in Y$), with no information provided as to which x_i matches which y_j.

Approach

- Learn two mappings:
  - *G: X → Y* and *F: Y → X*
- D_X and D_Y encourage generated outputs to be indistinguishable from target domain

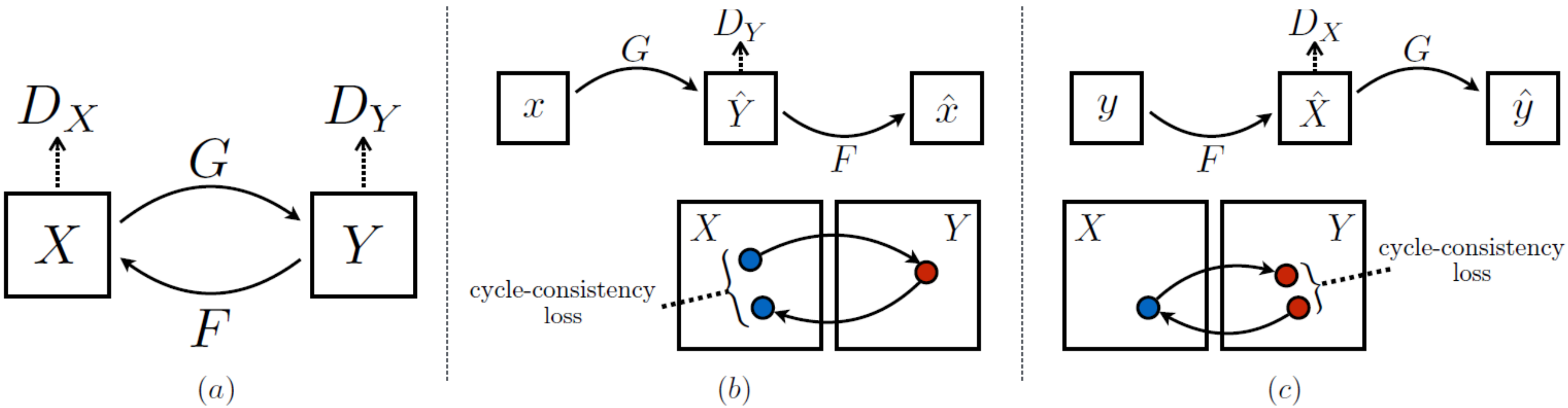


Figure 3: (a) Our model contains two mapping functions $G : X \rightarrow Y$ and $F : Y \rightarrow X$, and associated adversarial discriminators D_Y and D_X. D_Y encourages G to translate X into outputs indistinguishable from domain Y, and vice versa for D_X and F. To further regularize the mappings, we introduce two *cycle consistency losses* that capture the intuition that if we translate from one domain to the other and back again we should arrive at where we started: (b) forward cycle-consistency loss: $x \rightarrow G(x) \rightarrow F(G(x)) \approx x$, and (c) backward cycle-consistency loss: $y \rightarrow F(y) \rightarrow G(F(y)) \approx y$

Loss Functions

- Adversarial Loss: Match distribution of generated images to data distribution in the target domain

$$\begin{aligned}\mathcal{L}_{\text{GAN}}(G, D_Y, X, Y) =& \mathbb{E}_{y \sim p_{\text{data}}(y)}[\log D_Y(y)] \\ &+ \mathbb{E}_{x \sim p_{\text{data}}(x)}[\log(1 - D_Y(G(x))]\end{aligned}$$

- Cycle Consistency Loss: Prevent the learned mappings *G* and *F* from contradicting each other

$$\begin{aligned}\mathcal{L}_{\text{cyc}}(G, F) =& \mathbb{E}_{x \sim p_{\text{data}}(x)}[\|F(G(x)) - x\|_1] \\ &+ \mathbb{E}_{y \sim p_{\text{data}}(y)}[\|G(F(y)) - y\|_1].\end{aligned}$$

Loss Functions

- Full Objective:

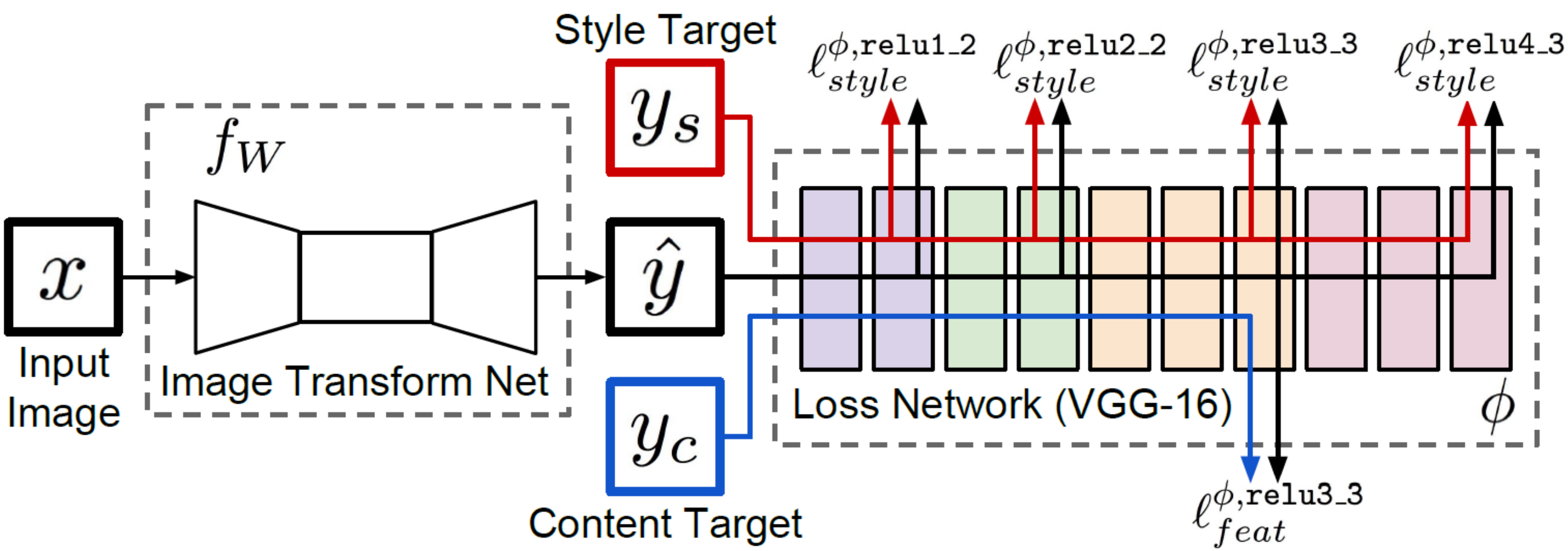
$$\mathcal{L}(G, F, D_X, D_Y) = \mathcal{L}_{\text{GAN}}(G, D_Y, X, Y) + \mathcal{L}_{\text{GAN}}(F, D_X, Y, X) + \lambda \mathcal{L}_{\text{cyc}}(G, F),$$

$$G^*, F^* = \arg \min_{G,F} \max_{D_x, D_Y} \mathcal{L}(G, F, D_X, D_Y)$$

- For training, $\lambda = 10$

Generators

- Generative network model from Johnson *et al.*
  - Two stride-2 convolutions, residual blocks, and two fractionally-strided convolutions with stride ½
  - 6 blocks for 128x128 images
  - 9 blocks for 256x256 and higher images



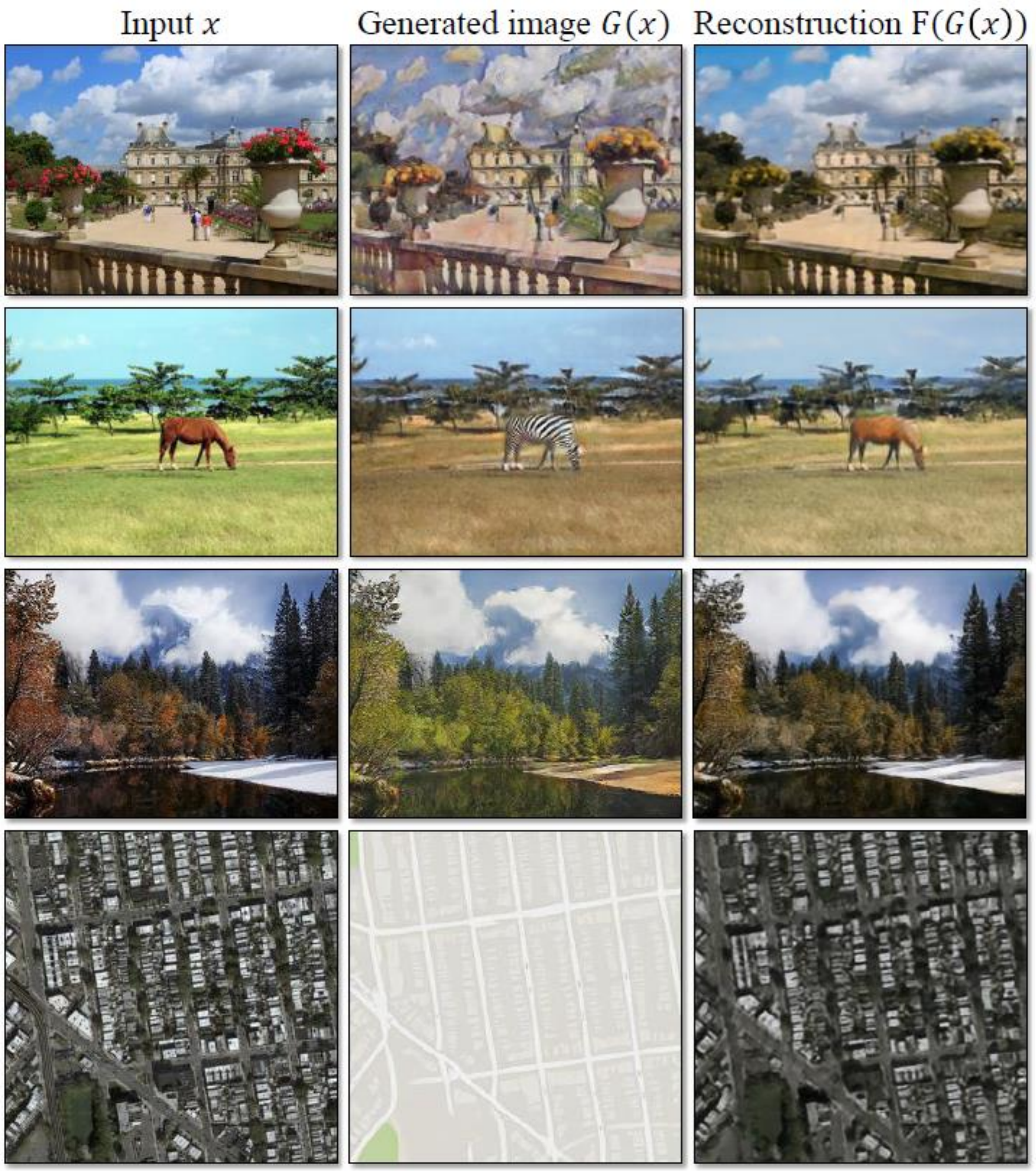
Johnson, J., Alahi, A., & Fei-Fei, L. "Perceptual losses for real-time style transfer and super-resolution." *ECCV 2016*.

Discriminators

- PatchGAN
  - Classify if each 70x70 patch in the image is real or fake
  - Run convolutionally across the image, averaging all responses to provide the final output
  - Fewer parameters
- Markov random field model, assuming independence between pixels separated by more than a patch diameter

Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. “Image-to-image translation with conditional adversarial networks.” *CVPR 2017.*

Cycle-Consistent Examples



Evaluation

- Cityscapes dataset: Semantic labels ←→Photo
- Google Maps: Map←→Aerial photo
- Amazon Mechanical Turk (AMT) assessment
    - Pick image they think is real
- Fully-convolutional network (FCN)
    - Predict label map for generated photo
- Per-pixel accuracy, per-class accuracy, and mean class intersection-over-union (IOU)

Baselines for Comparison

CoGAN [30] This method learns one GAN generator for domain X and one for domain Y, with tied weights on the first few layers for shared latent representation. Translation from X to Y can be achieved by finding a latent representation that generates image X and then rendering this latent representation into style Y.

SimGAN [45] Like our method, Shrivastava et al.[45] uses an adversarial loss to train a translation from X to Y. The regularization term $\|X - G(X)\|_1$ was used to penalize making large changes at pixel level.

Feature loss + GAN We also test a variant of SimGAN [45] where the L1 loss is computed over deep image features using a pretrained network (VGG-16 `relu4_2` [46]), rather than over RGB pixel values. Computing distances in deep feature space, like this, is also sometimes referred to as using a "perceptual loss" [7, 22].

BiGAN/ALI [8, 6] Unconditional GANs [15] learn a generator $G : Z \rightarrow X$, that maps random noise Z to images X. The BiGAN [8] and ALI [6] propose to also learn the inverse mapping function $F : X \rightarrow Z$. Though they were originally designed for mapping a latent vector z to an image x, we implemented the same objective for mapping a source image x to a target image y.

pix2pix [21] We also compare against pix2pix [21], which is trained on paired data, to see how close we can get to this "upper bound" without using any paired training data.

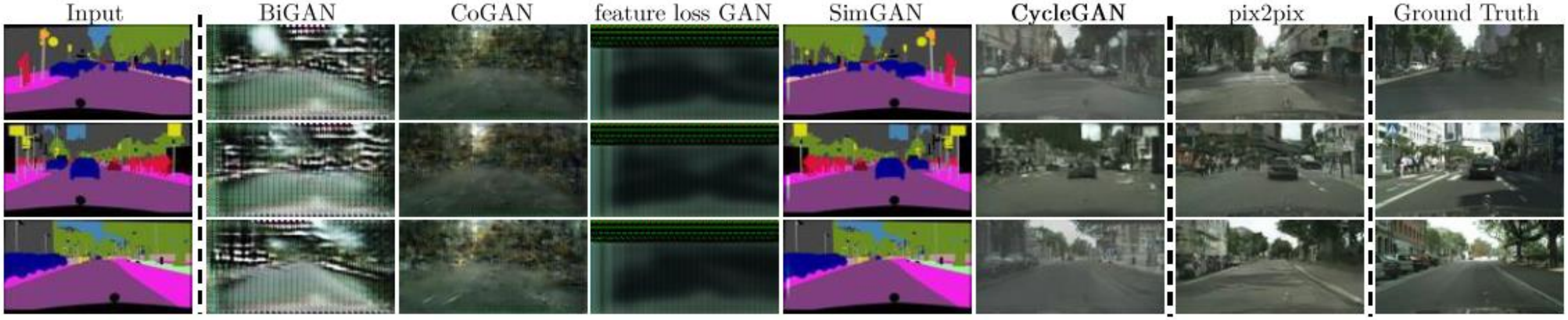


Figure 5: Different methods for mapping labels↔photos trained on Cityscapes images. From left to right: input, BiGAN/ALI [6, 8], CoGAN [30], feature loss + GAN, SimGAN [45], CycleGAN (ours), pix2pix [21] trained on paired data, and ground truth.

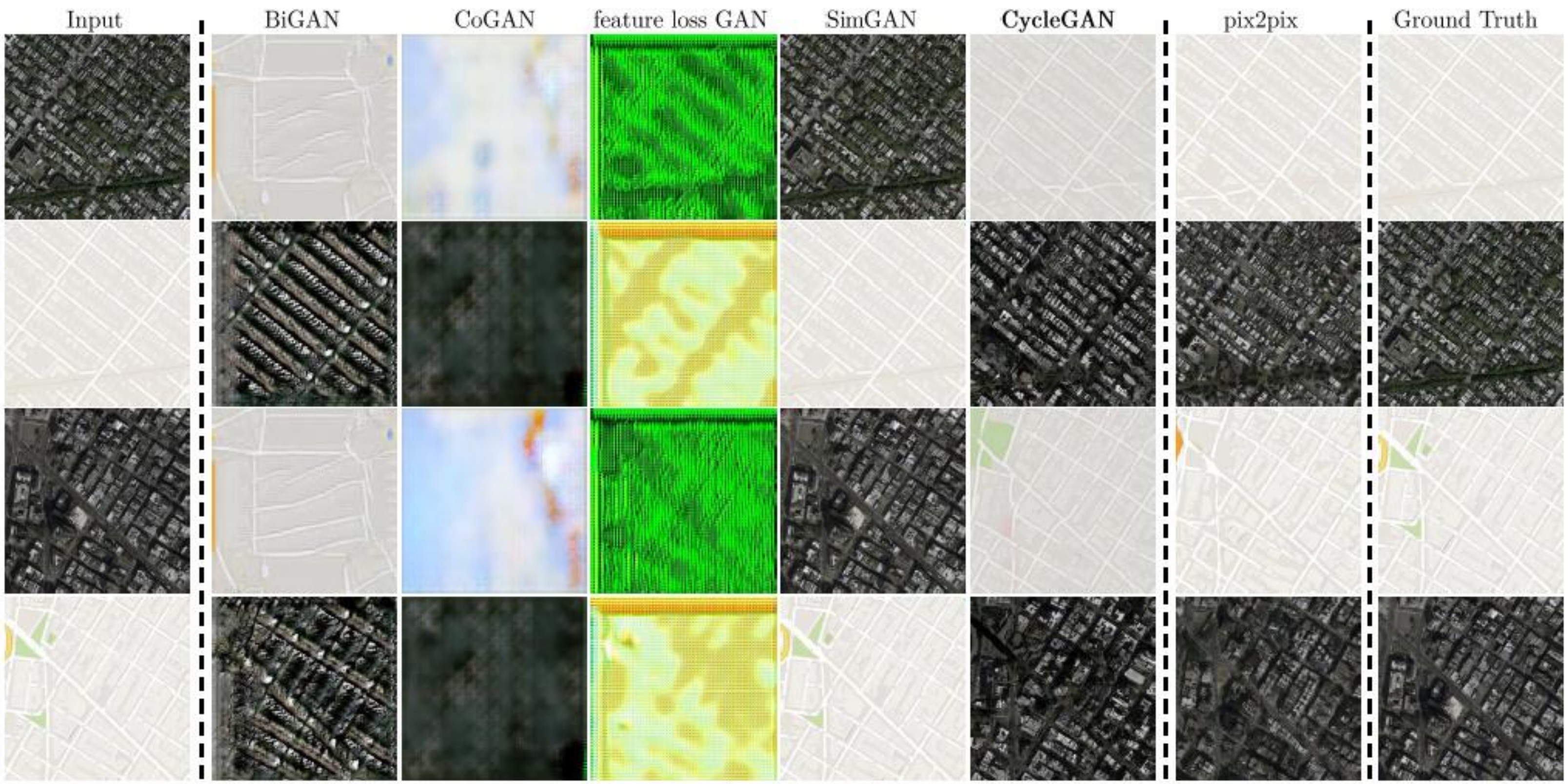


Figure 6: Different methods for mapping aerial photos↔maps on Google Maps. From left to right: input, BiGAN/ALI [6, 8], CoGAN [30], feature loss + GAN, SimGAN [45], CycleGAN (ours), pix2pix [21] trained on paired data, and ground truth.

Loss	Map → Photo % Turkers labeled *real*	Photo → Map % Turkers labeled *real*
CoGAN [30]	$0.6\% \pm 0.5\%$	$0.9\% \pm 0.5\%$
BiGAN/ALI [8, 6]	$2.1\% \pm 1.0\%$	$1.9\% \pm 0.9\%$
SimGAN [45]	$0.7\% \pm 0.5\%$	$2.6\% \pm 1.1\%$
Feature loss + GAN	$1.2\% \pm 0.6\%$	$0.3\% \pm 0.2\%$
CycleGAN (ours)	**$26.8\% \pm 2.8\%$**	**$23.2\% \pm 3.4\%$**

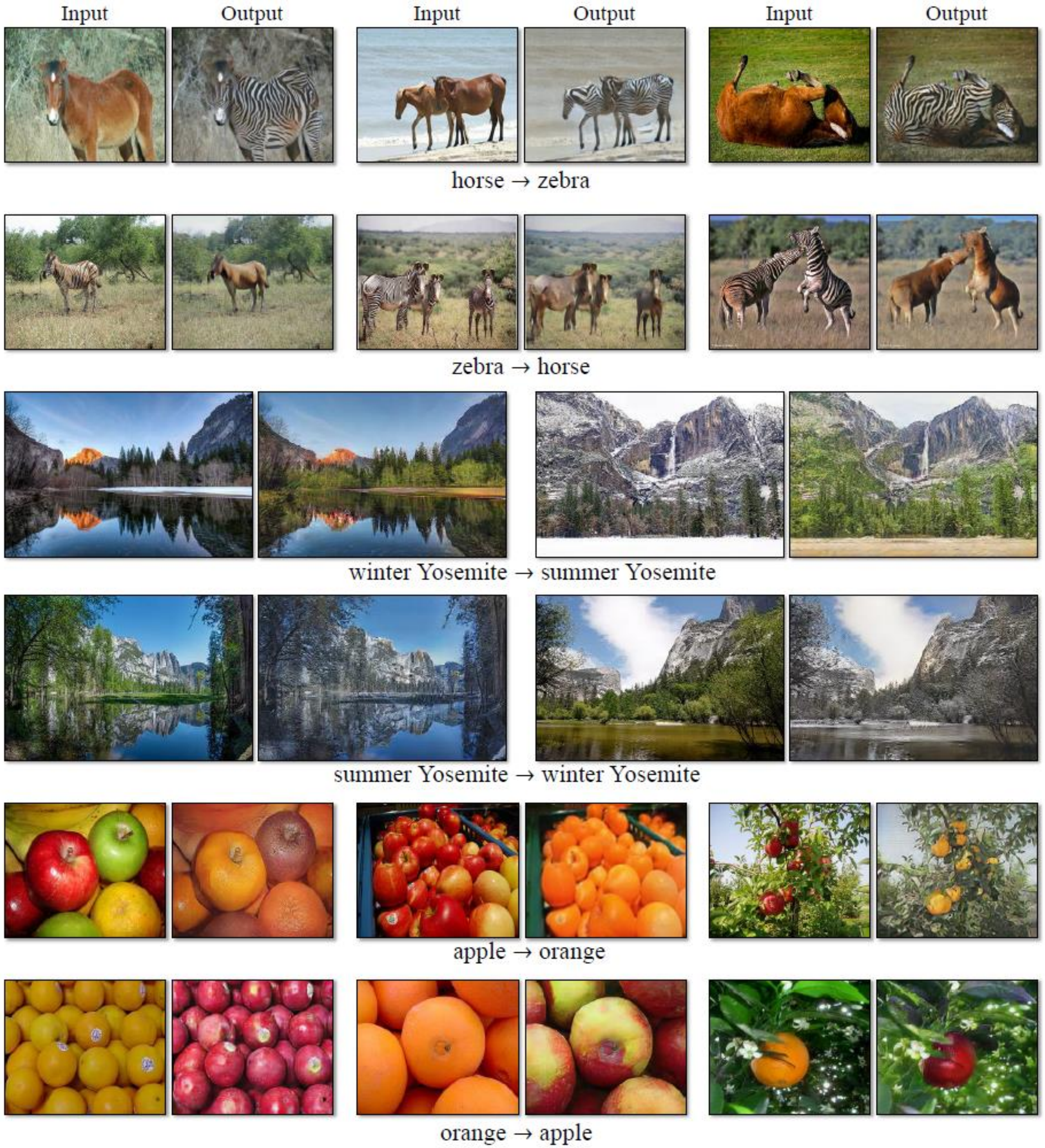
Table 1: AMT "real vs fake" test on maps↔aerial photos at 256×256 resolution.

Loss	Per-pixel acc.	Per-class acc.	Class IOU
CoGAN [30]	0.40	0.10	0.06
BiGAN/ALI [8, 6]	0.19	0.06	0.02
SimGAN [45]	0.20	0.10	0.04
Feature loss + GAN	0.06	0.04	0.01
CycleGAN (ours)	**0.52**	**0.17**	**0.11**
pix2pix [21]	0.71	0.25	0.18

Table 2: FCN-scores for different methods, evaluated on Cityscapes labels→photo.

Art Style Transfer

Transfiguration



Other Applications: Face ←→ Ramen

Ramen Input

Face Input

Paper Conclusions

- Compelling results on translation tasks that involve color and texture changes
- Tasks that require geometric changes are less successful
- Generator architecture tailored for appearance changes
- May need to incorporate weak semantic supervision

cat → dog

Horse → Zebra