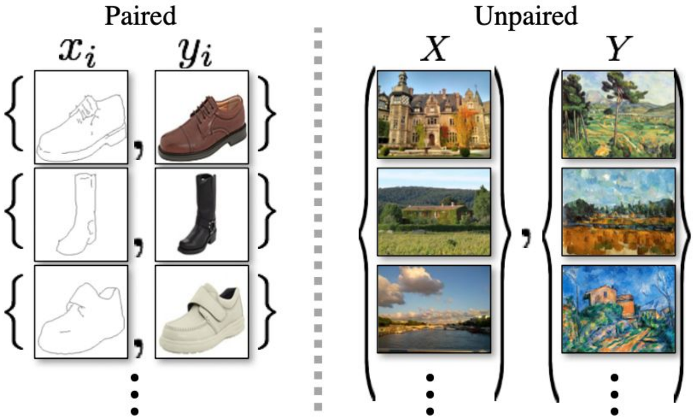
Midterm 4
Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks

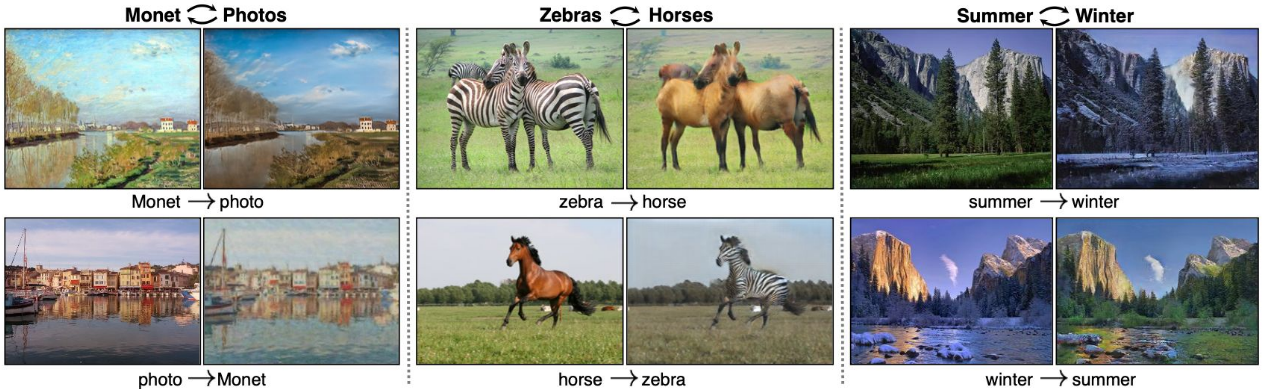
Lorenzo Leuzzi

Intelligent System for Pattern Recognition 2023

Introduction to the problem

Traditionally, image-to-image translation tasks require **paired data**, where each image in one domain has a corresponding image in the other domain. However, collecting such paired data can be time-consuming and expensive. **CycleGAN** allows for learning mappings between two domains **without** direct **correspondences**.



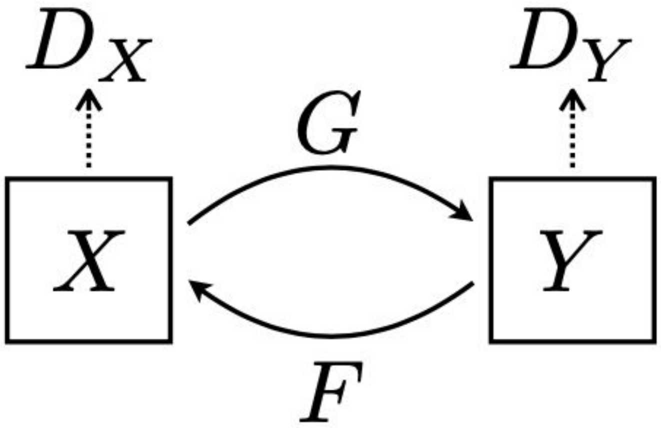


Model Description

The model includes:

- two **generators** :
  G *: X → Y*, tries to generate images *G(x)* that look similar to images from domain *Y*;
  F *: Y → X*, tries to generate images *F(y)* that look similar to images from domain *X*.

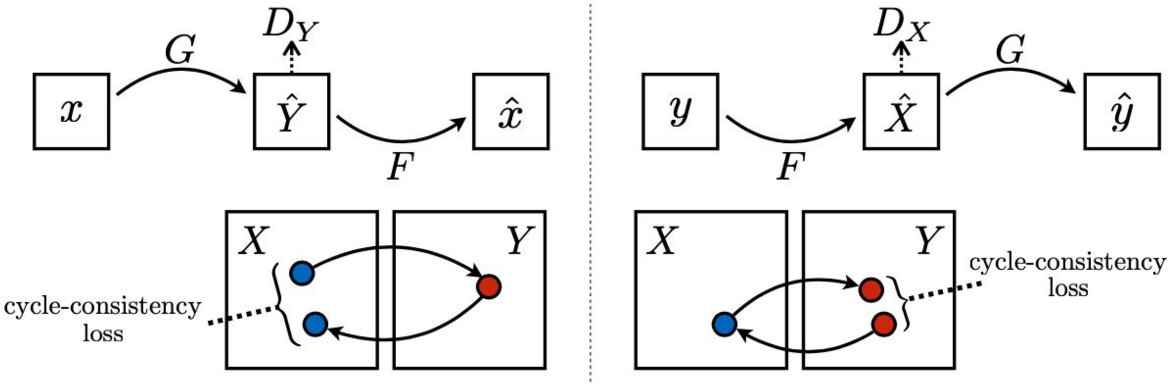
- two **discriminators**:
  D_X aims to distinguish between images from the real distribution *{x}* and translated images *{F(y)}*;
  D_Y aims to discriminate between *{y}* and *{G(x)}*.



Cycle consistency

Problem: with large enough capacity, a network can map the same set of input images to any random permutation of images in the target domain, where any of the learned mappings can induce an output distribution that matches the target distribution. Thus, **adversarial losses** alone is **not sufficient**.

We want that if we take an image from domain X, convert it to domain Y, and then convert it back to domain X, the resulting image should be similar to the original image. This is obtained by training both the mapping G and F simultaneously, and adding a **cycle consistency loss** that encourages $F(G(x)) \approx x$ and $G(F(y)) \approx y$.



The catch

Adversarial Loss
G aims to minimize this objective s.t. *D(G(x)) → 1,* against an adversary D that tries to maximize it s.t. *D(real) → 1* and *D(fake) → 0*.

Similarly for F.

distribution of the data from target domain

D output for real data

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

distribution of the data from source domain

D output for generated (fake) data

$$\mathcal{L}_{\text{GAN}}(F, D_X, Y, X)$$

Cycle Consistency Loss
For each image *x* from domain *X*, the image translation cycle should be able to bring *x* back to the original image: *x → G(x) → F (G(x)) ≈ x*.

Similarly for each image *y*.

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

difference between the reconstructed image and the real image (after a full cycle)

Full Objective

Putting all together

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

adversarial losses for both mapping $G: X \to Y$ with D_Y and the mapping $F: Y \to X$ with D_X

cycle consistency loss where λ controls the relative importance of the two objectives

We aim to solve the optimization problem

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

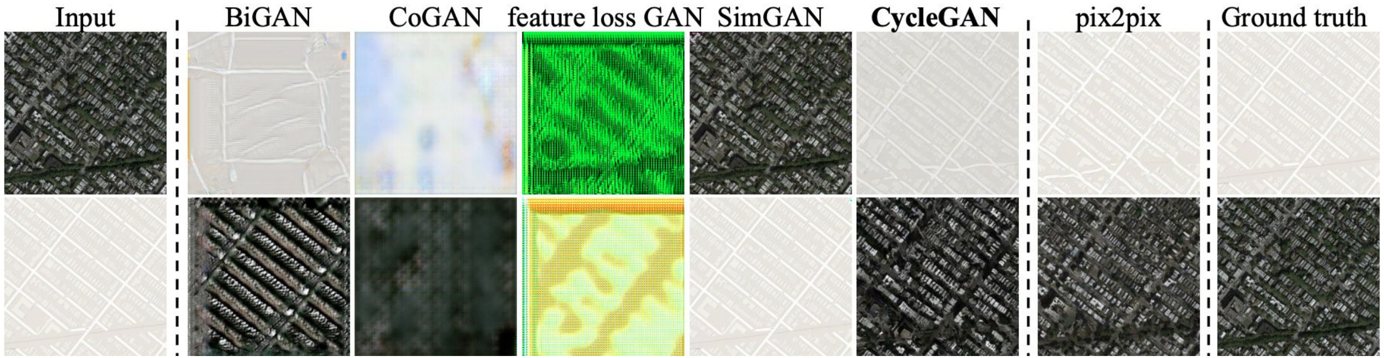
Results

Assessment method: participants were shown a sequence of pairs of images, one a real photo or map and one fake (generated by our algorithm or a baseline), and asked to click on the image they thought was real.

These results not only demonstrate that the images generated by CycleGAN were **highly realistic** compared to those generated by other unsupervised models, but also indicate that CycleGAN was successful in deceiving human observers approximately ¼ of the time.

Loss	**Map → Photo** % Turkers labeled *real*	**Photo → Map** % Turkers labeled *real*
CoGAN [32]	$0.6\% \pm 0.5\%$	$0.9\% \pm 0.5\%$
BiGAN/ALI [9, 7]	$2.1\% \pm 1.0\%$	$1.9\% \pm 0.9\%$
SimGAN [46]	$0.7\% \pm 0.5\%$	$2.6\% \pm 1.1\%$
Feature loss + GAN	$1.2\% \pm 0.6\%$	$0.3\% \pm 0.2\%$
CycleGAN (ours)	$\mathbf{26.8\% \pm 2.8\%}$	$\mathbf{23.2\% \pm 3.4\%}$

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



Comments

On translation tasks that involve color and texture changes the method often succeeds producing very credible images. The **successful** application of the model are various: photo enhancement, style transfer (e.g. season change, artwork style transfer), object transfiguration (e.g. horse to zebra, apple to orange), etc.

Weak point:
- tasks that require geometric changes, CycleGAN can only make minimal changes to the input (e.g. dog to cat);
- changes in distribution characteristics (training set vs test set) can result in failure cases.

The gap between the results achievable with paired training data (pix2pix) and those achieved by our unpaired method is noticeable.
Integrating weak or semi-supervised data may lead to substantially more powerful translators.

In conclusion, **CycleGAN framework is a powerful tool for image-to-image translations that overcomes the need for paired training data and this paper *pushed the boundaries of what is possible in this unsupervised setting.***