II SIMPOSIO PERUANO DE DEEP LEARNING

Del 8 al 10 de Enero 2020

Universidad Tecnológica del Perú

GANs para la traducción de imágenes

Pablo Fonseca, PUCP / UPC

Presentación para el **II Simposio Peruano de Deep Learning**, 8 al 10 de Enero 2020
Arequipa-Perú / Charla Online

Outline

01. Presentación
02. Introducción: Machine/Deep Learning
03. Redes GANs
04. Traducción de Imágenes
05. Traducción de Imágenes: Pix2Pix
06. Traducción de Imágenes: CycleGAN
07. Traducción de Imágenes: Pix2PixHD

Presentación

- En esta charla discutiremos algunas ideas de traducción de imágenes en el framework de las redes adversariales (GANs)
- En particular revisaremos 3 métodos y sus ideas principales:
  - **Pix2Pix** (mapear imágen a imágen, más allá de la reconstrucción L1/L2)
  - **CycleGAN** (aprovechar la información de los conjuntos cuando no tenemos supervisión “explícita” por imagen)
  - **Pix2PixHD** (Perceptual Loss y algunas ideas de cómo lograr alta resolución)
- Nota: Estos no son los únicos métodos de traducción de imágenes

01

Introducción: Machine Learning y Deep Learning

¿Qué es Machine Learning?

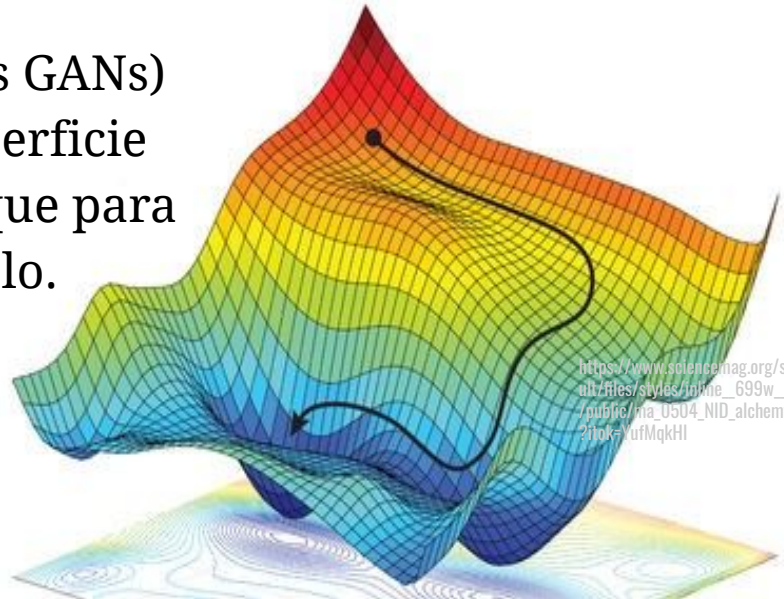
“Se puede decir que un programa aprende de una experiencia **E** con respecto a cierto tipo de tareas **T** con una medida de performance **P** si la perfomance en las tareas del tipo **T**, medidas por **P**, mejora con la experiencia **E**.”

Tom Mitchell

Aprendizaje y Optimización

Se puede transformar un problema de aprendizaje en uno de optimización

Cuando convertimos un juego (como los GANs) en un problema de optimización, la superficie del error puede ser “más complicada” que para un problema de clasificación por ejemplo.



Redes GANs

03

GANs

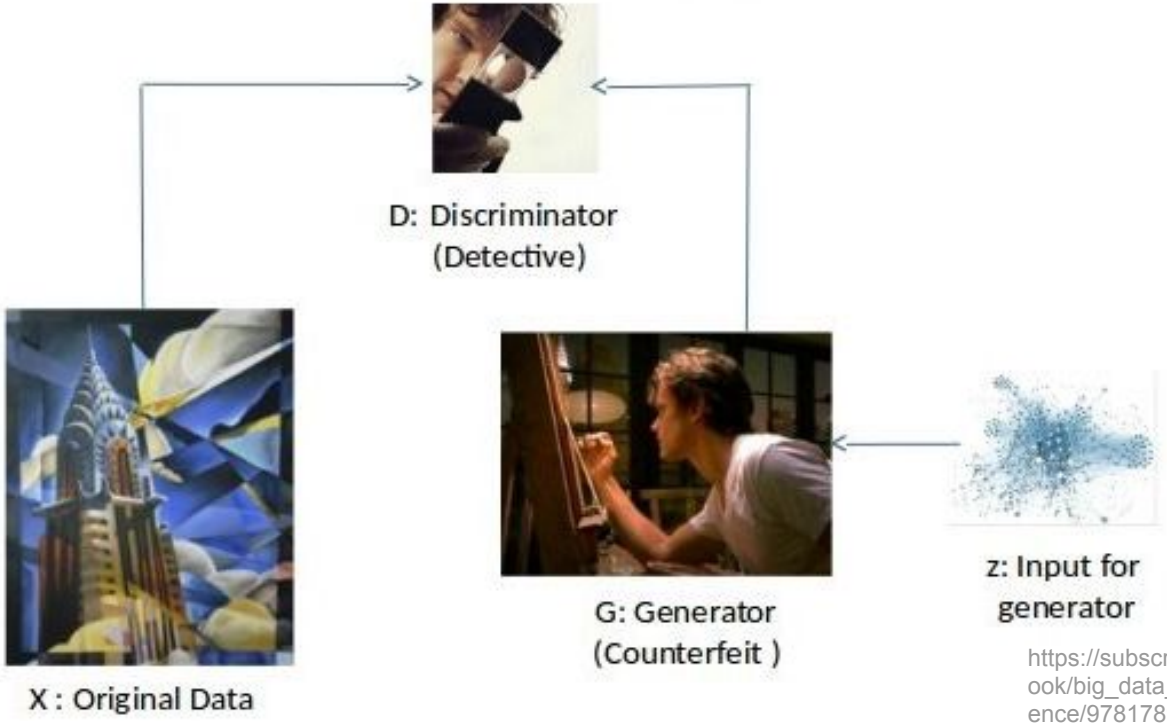
X : Original Data

G: Generator
(Counterfeit)

z: Input for
generator

https://subscription.packtpub.com/book/big_data_and_business_intelligence/9781789538205/2/ch02lvl1sec16/gans-building-blocks

GANs



https://subscription.packtpub.com/book/big_data_and_business_intelligence/9781789538205/2/ch02lvl1sec16/gans-building-blocks

Discriminador

Es una función de pérdida “adaptativa”, que depende de los datos

GANs: Esta persona no existe

Ver más en: **https://thispersondoesnotexist.com/**

GANs

Primera obra de arte (I.A.) vendida por $400K usando GANs

Is artificial intelligence set to become art's next medium?

AI artwork sells for $432,500 — nearly 45 times its high estimate — as Christie's becomes the first auction house to offer a work of art created by an algorithm

RINTS |

https://www.christies.com/features/A-collaboration-between-two-artists-one-human-one-a-machine-9332-1.aspx

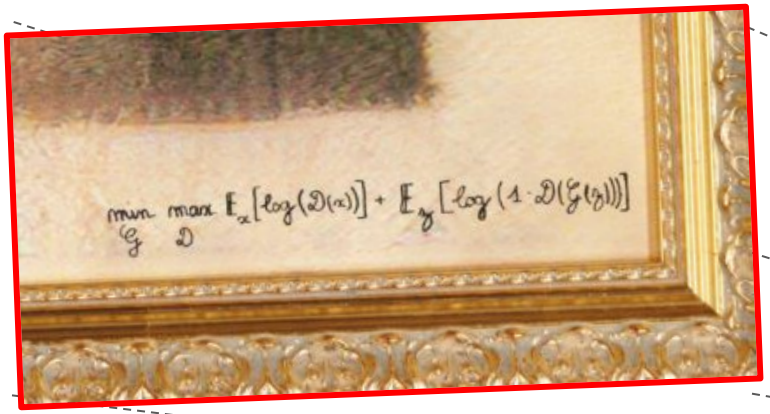
GANs

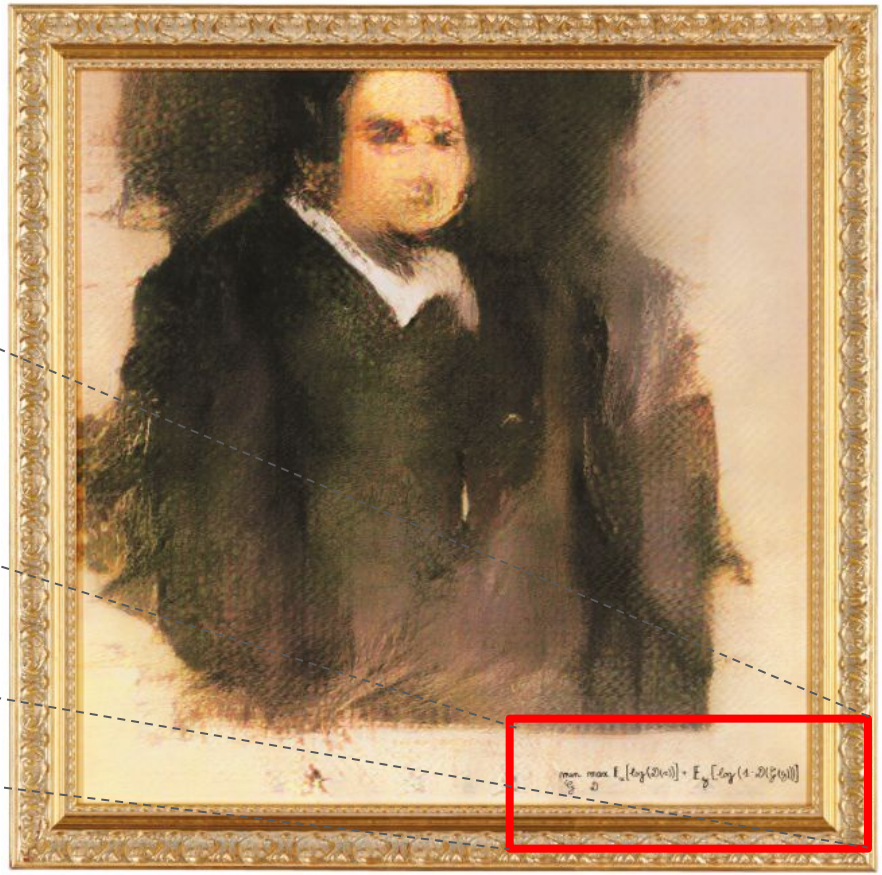
Esta es la pintura completa

https://www.christies.com/features/A-collaboration-between-two-artists-one-human-one-a-machine-9332-1.aspx

GANs

Función de pérdida





https://www.christies.com/features/A-collaboration-between-two-artists-one-human-one-a-machine-9332-1.aspx

Traducción de Imágenes

04

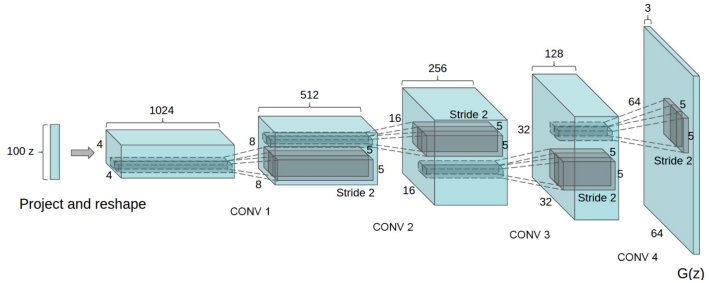
Línea de Tiempo

Paper original de GAN

Goodfellow et al.

Línea de Tiempo

DC GAN
Radford et al
2015



2014 2015 2016 2017 2018 2019 2020

Paper original de GAN

Goodfellow et al.
2014

Línea de Tiempo

DC GAN
Radford et al
2015

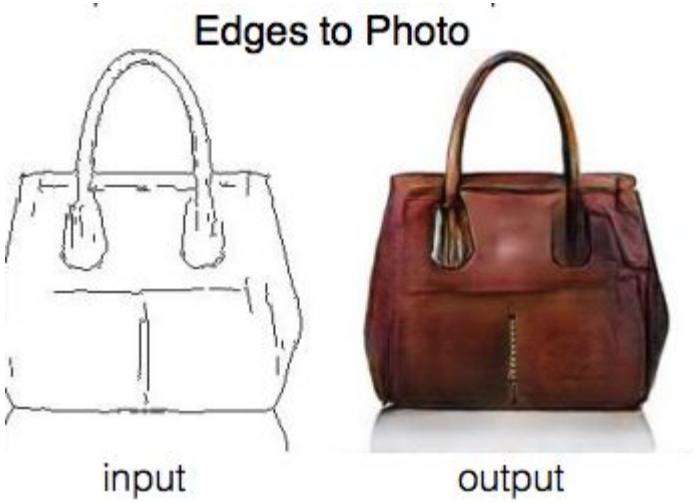
2014 2015 2016 2017 2018 2019 2020

Paper original de GAN

Goodfellow et al. 2014

Pix2Pix

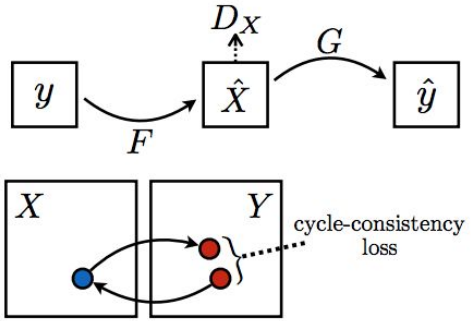
Isola et al. 2016



Línea de Tiempo

DC GAN
Radford et al
2015

CycleGAN
Zhu et al. (2017)



Paper original de GAN

Goodfellow et al. 2014

Pix2Pix

Isola et al. 2016

Línea de Tiempo

DC GAN
Radford et al
2015

CycleGAN
Zhu et al. (2017)

2014 2015 2016 2017 2018 2019 2020

Paper original de GAN

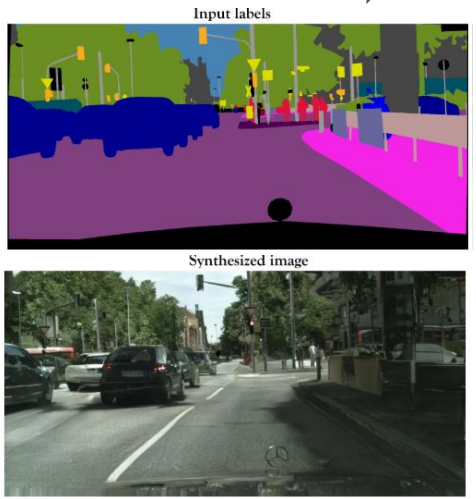
Goodfellow et al. 2014

Pix2Pix

Isola et al. 2016

Pix2Pix HD

Wang et al. 2018



Línea de Tiempo

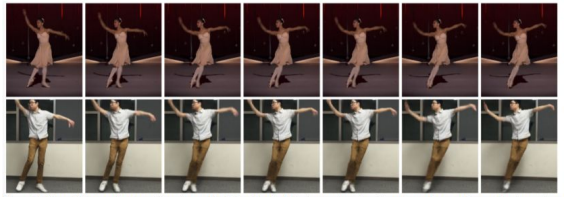


Figure 1: **"Do as I Do" motion transfer:** given a YouTube clip of a ballerina (top), and a video of a graduate student performing various motions, our method transfers the ballerina's performance onto the student (bottom). Video: https://youtu.be/mSaIrz8lM1U

DC GAN
Radford et al
2015

CycleGAN
Zhu et al. (2017)

Everybody Dance Now

Chan et al. 2018

2014 2015 2016 2017 2018 2019 2020

Paper original de GAN

Goodfellow et al. 2014

Pix2Pix

Isola et al. 2016

Pix2Pix HD

Wang et al. 2018

Traducción de Imágenes: Pix2Pix

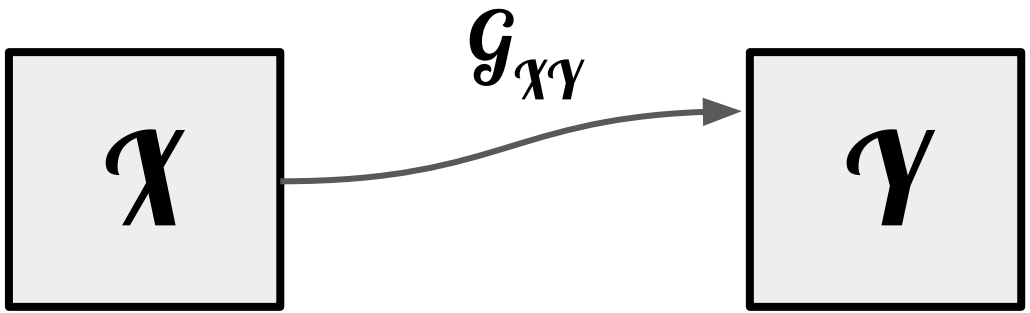
05

Image-to-Image Translation with Conditional Adversarial Nets

Phillip Isola Jun-Yan Zhu Tinghui Zhou Alexei A. Efros

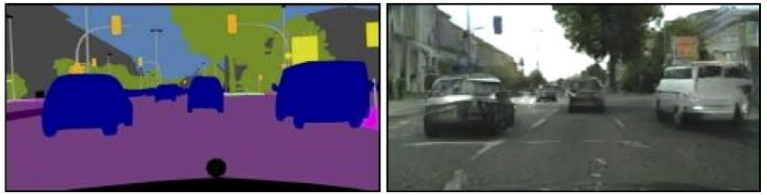
University of California, Berkeley
In CVPR 2017

Traducción de Imágenes



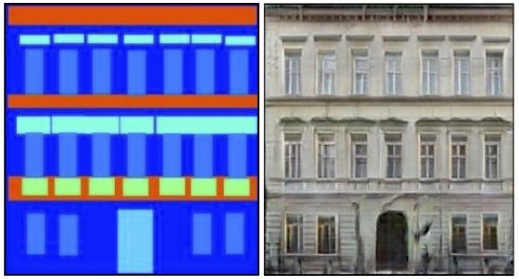
Traducción de Imágenes

Labels to Street Scene



input output

Labels to Facade



input output

BW to Color

input output

Aerial to Map

input output

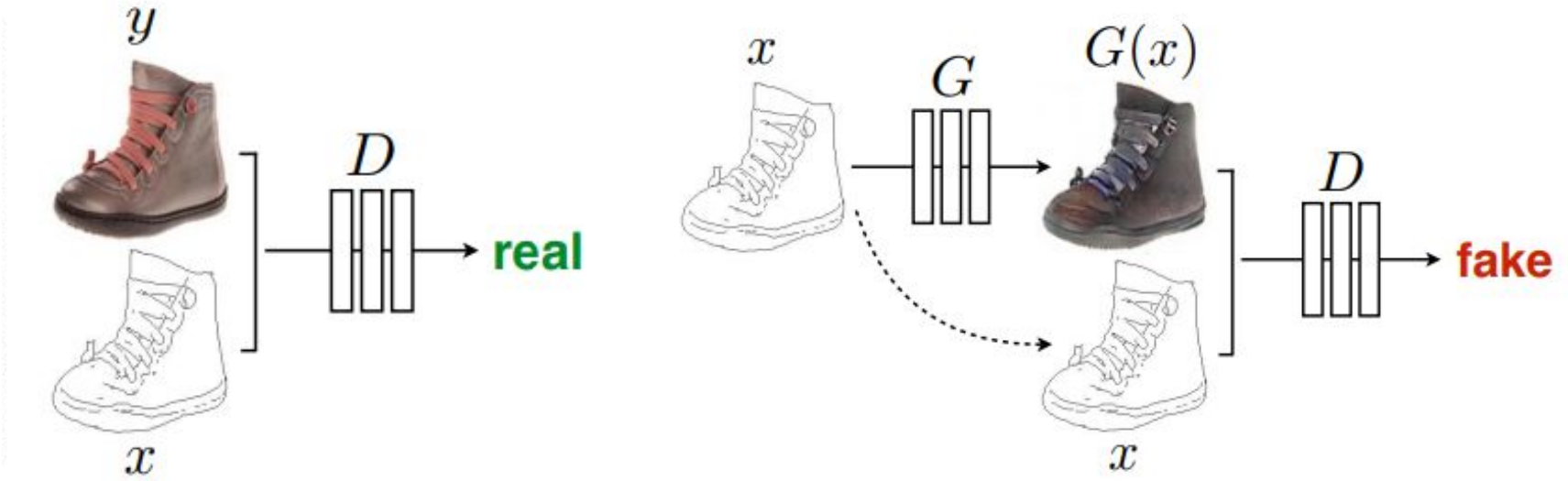
Day to Night

input output

Edges to Photo

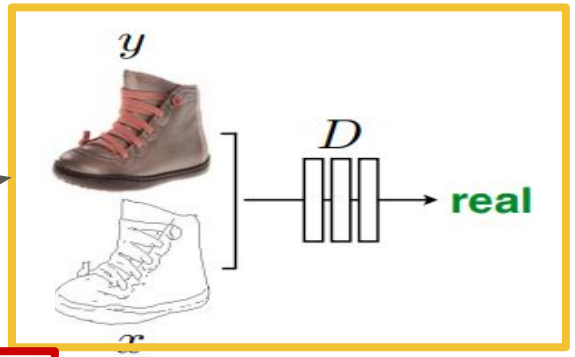
input output

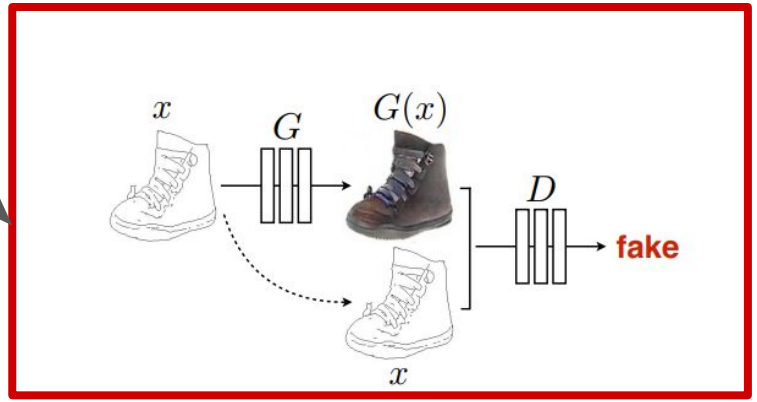
Métodos



Métodos: Función de pérdida

$$\mathcal{L}_{cGAN}(G, D) = \mathbb{E}_{x,y}[\log D(x, y)] + \mathbb{E}_{x,z}[\log(1 - D(x, G(x, z))]$$





Esta parte del objetivo está relacionada al generador **G** produciendo imágenes que engañen a **D**

Usando L1 para más nitidez

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y,z}[\|y - G(x,z)\|_1].$$

Usando L1 para más nitidez

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y,z}[\|y - G(x,z)\|_1].$$

$$G^* = \arg\min_G \max_D \mathcal{L}_{cGAN}(G,D) + \lambda \mathcal{L}_{L1}(G).$$

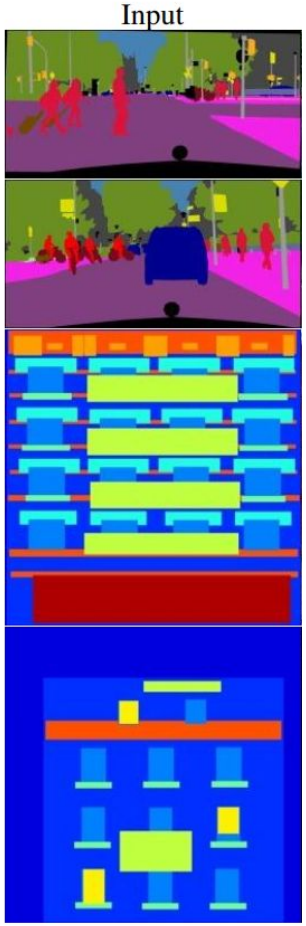
Usando L1 para más nitidez

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y,z}[\|y - G(x,z)\|_1].$$

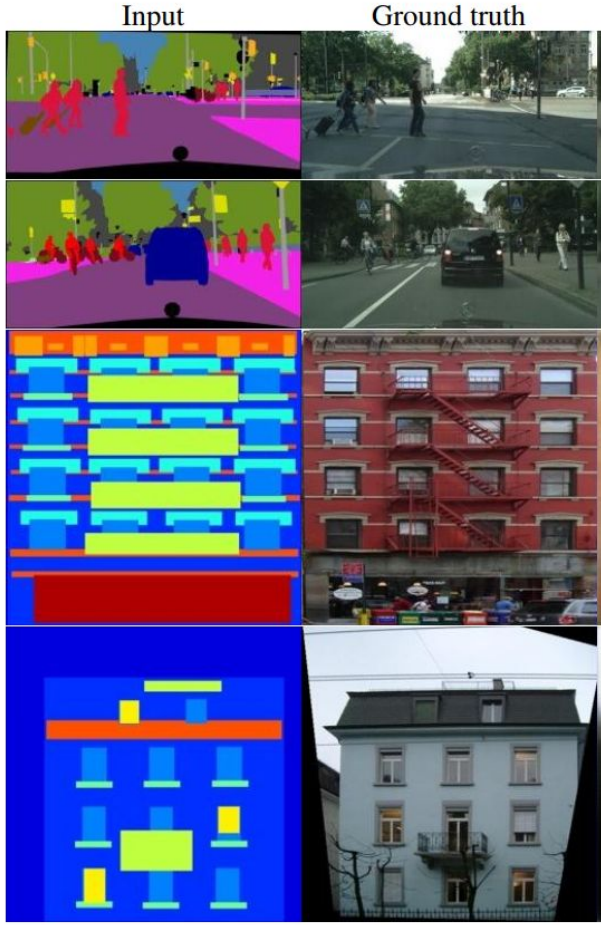
$$G^* = \arg\min_G \max_D \mathcal{L}_{cGAN}(G,D) + \lambda\mathcal{L}_{L1}(G).$$

Componente L1

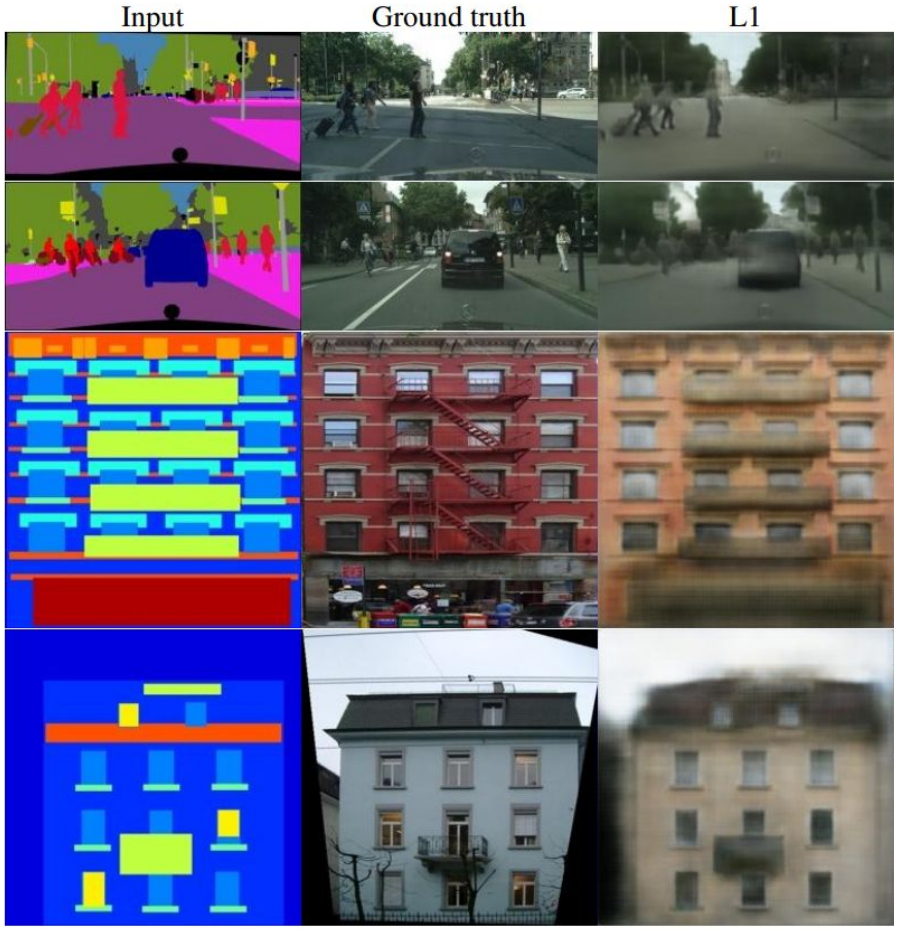
Usando L1 para más nitidez



Usando L1 para más nitidez



Usando L1 para más nitidez

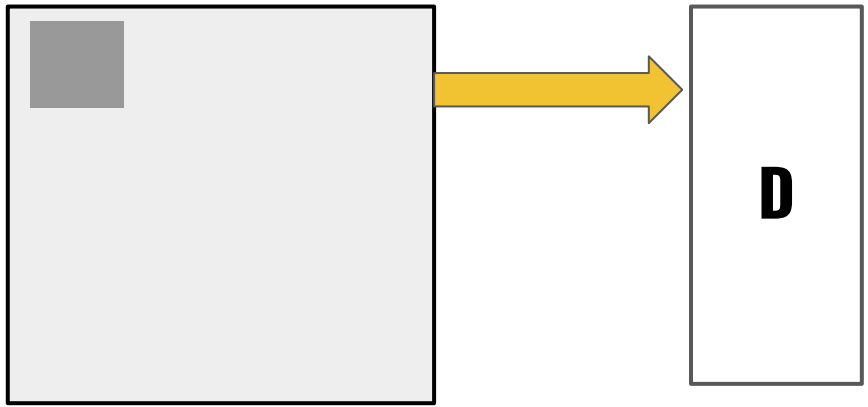


Usando L1 para más nitidez

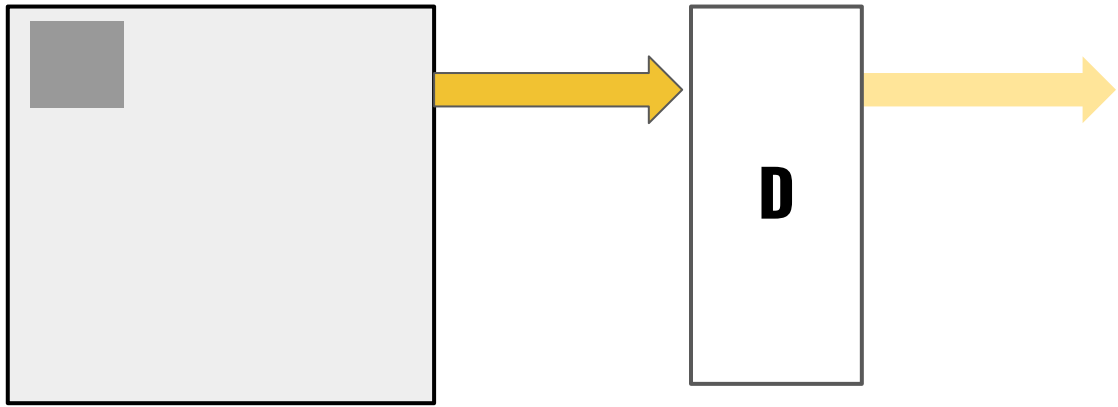
Usando L1 para más nitidez

Discriminador (textura y alta frecuencia)

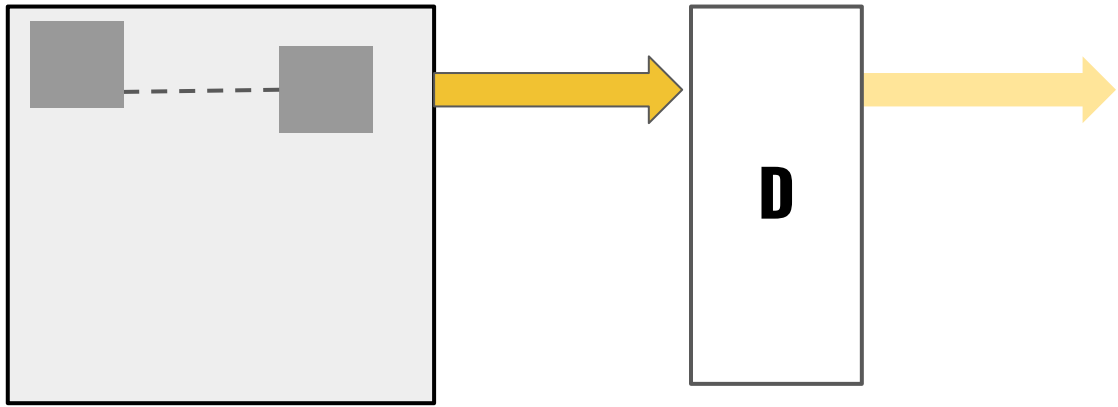
Discriminador (textura y alta frecuencia)



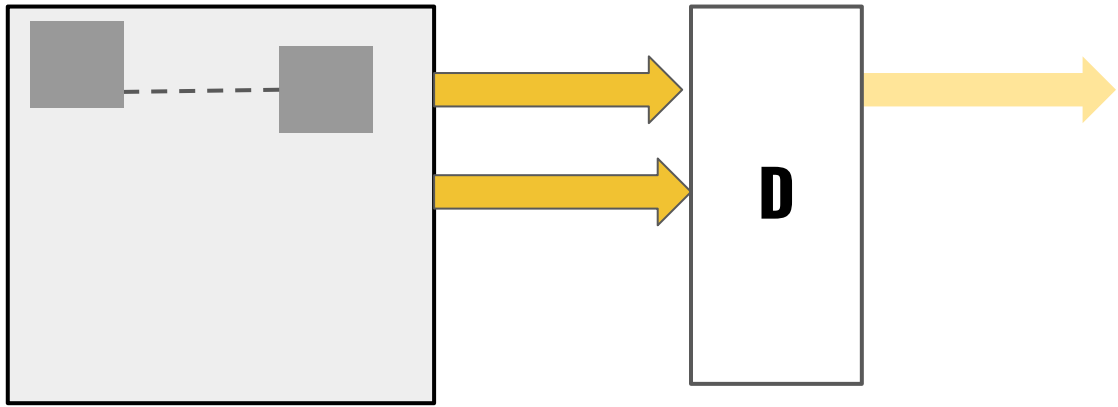
Discriminador (textura y alta frecuencia)



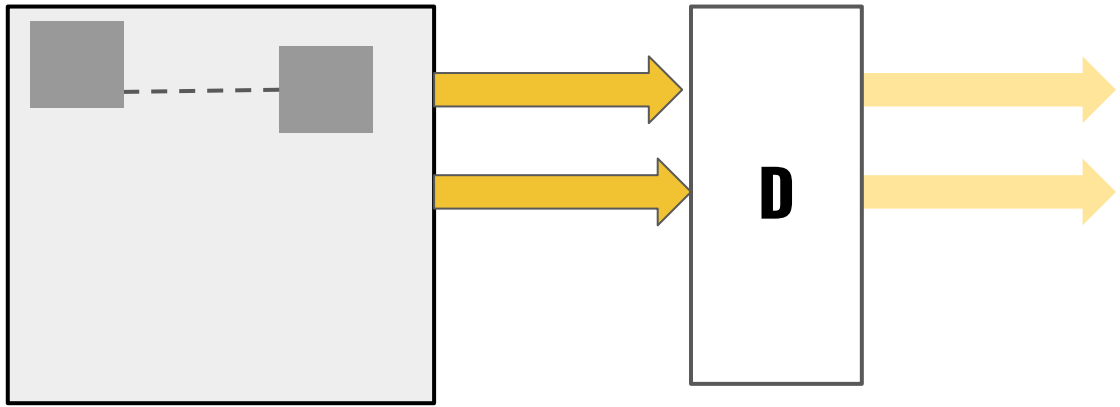
Discriminador (textura y alta frecuencia)



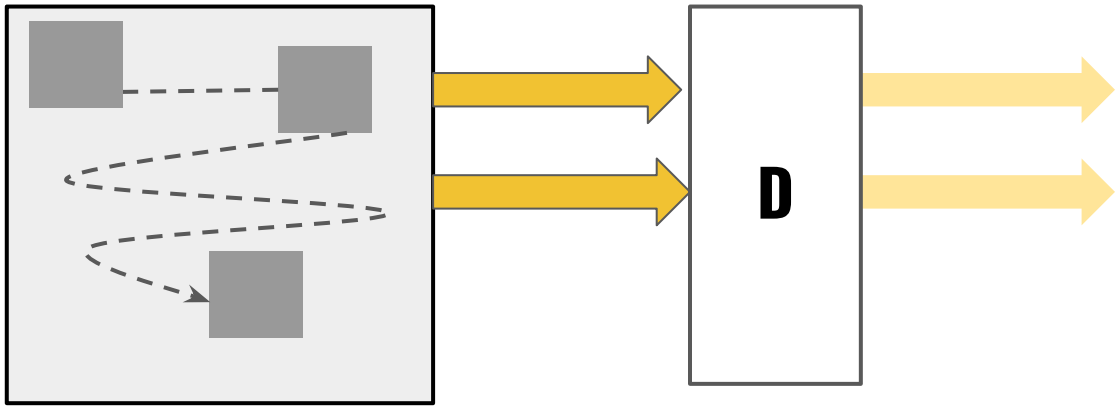
Discriminador (textura y alta frecuencia)



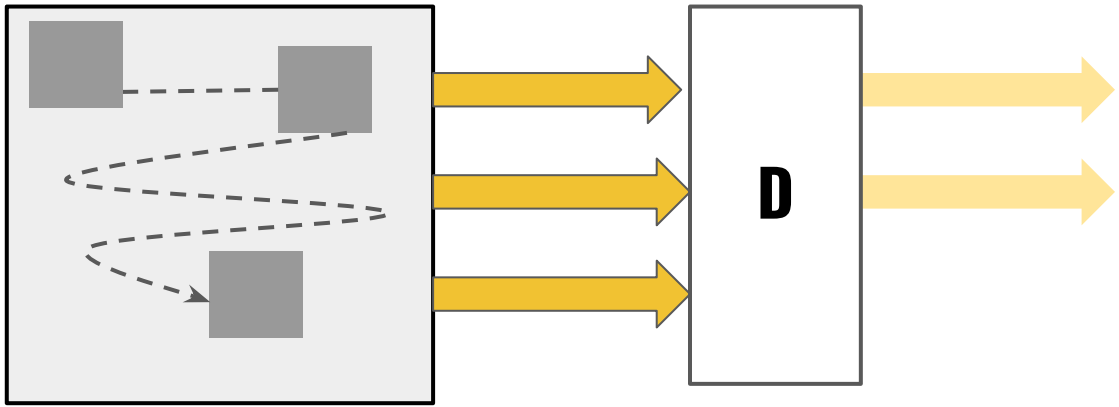
Discriminador (textura y alta frecuencia)



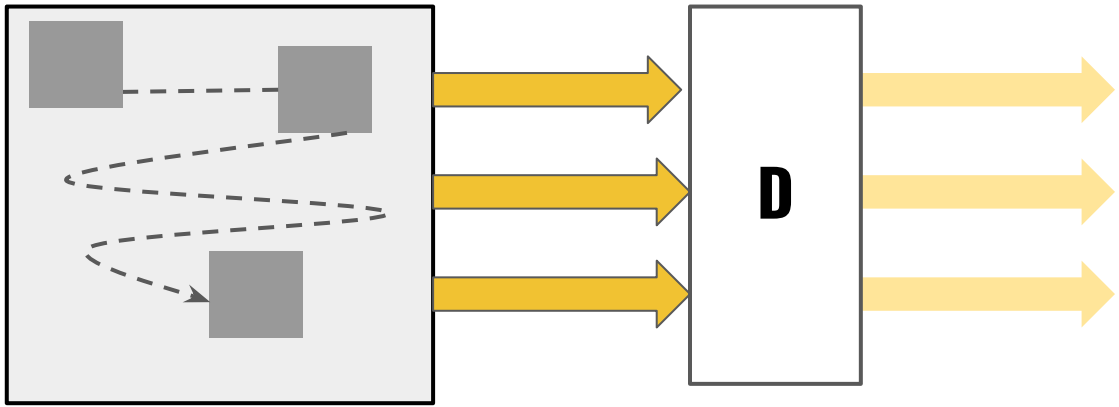
Discriminador (textura y alta frecuencia)



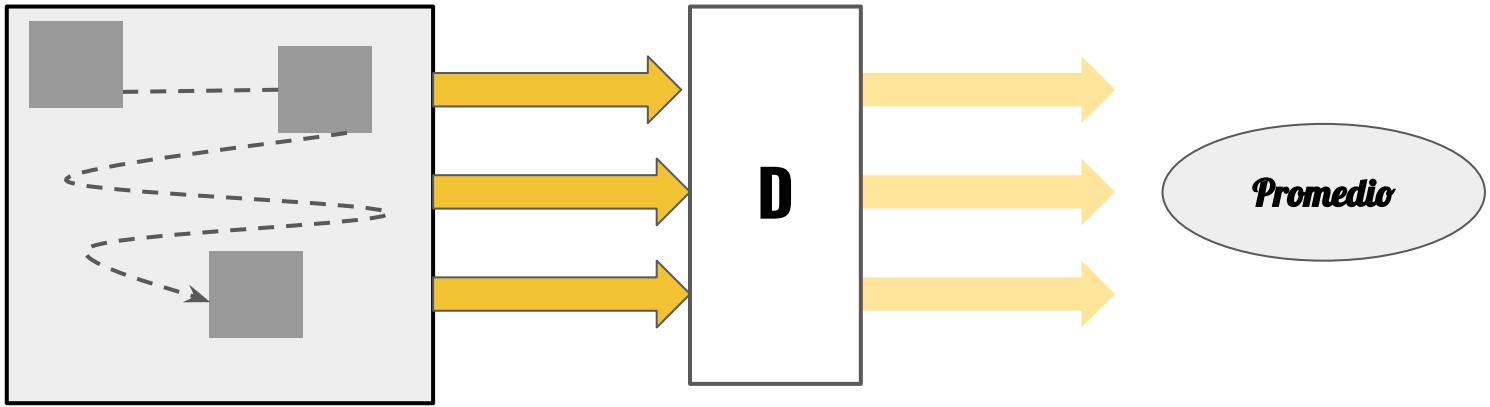
Discriminador (textura y alta frecuencia)



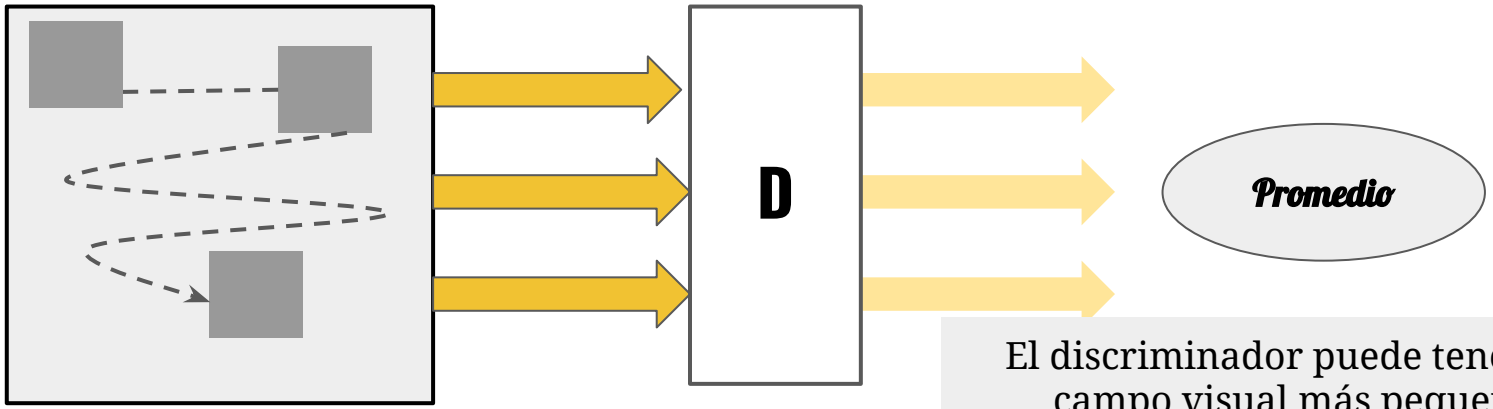
Discriminador (textura y alta frecuencia)



Discriminador (textura y alta frecuencia)



Discriminador (textura y alta frecuencia)

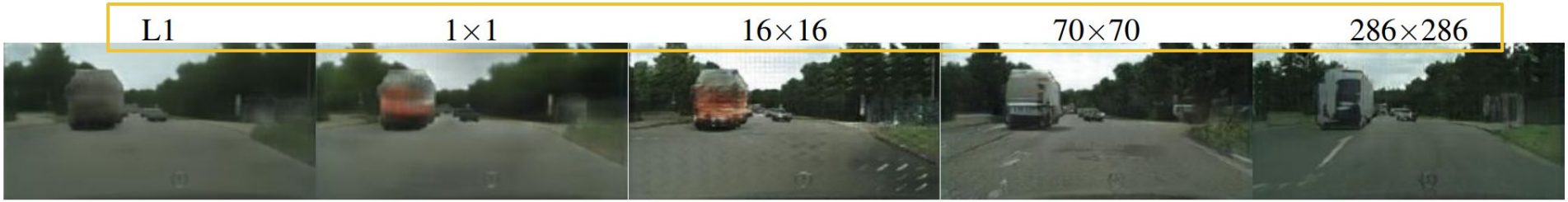


El discriminador puede tener un campo visual más pequeño

Cambiando el tamaño del “patch”

Cambiando el tamaño del “patch”

El tamaño del campo visual de **D**



Cambiando el tamaño del “patch”

Sin PatchGan

Cambiando el tamaño del “patch”

Sin PatchGan

Tamaño completo

Cambiando el tamaño del “patch”

No son muy diferentes

Sin PatchGan

Tamaño completo

Métodos: L1 Loss para más nitidez

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y,z}[\|y - G(x,z)\|_1].$$

Métodos: L1 Loss para más nitidez

Comparado a L2

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y,z}[\|y - G(x,z)\|_1].$$

Métodos: L1 Loss para más nitidez

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y,z}[\|y - G(x,z)\|_1].$$

Una de las restricciones es que la imagen resultante esté cerca del GT en términos de la distancia L1

Métodos: L1 Loss para más nitidez

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y,z}[\|y - G(x,z)\|_1].$$

$$G^* = \arg\min_G \max_D \mathcal{L}_{cGAN}(G,D) + \lambda \mathcal{L}_{L1}(G).$$

Métodos: L1 Loss para más nitidez

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y,z}[\|y - G(x,z)\|_1].$$

$$G^* = \arg\min_G \max_D \mathcal{L}_{cGAN}(G,D) + \lambda\mathcal{L}_{L1}(G).$$

L1 (lambda=100.0)

Traducción de Imágenes: CycleGAN

06

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

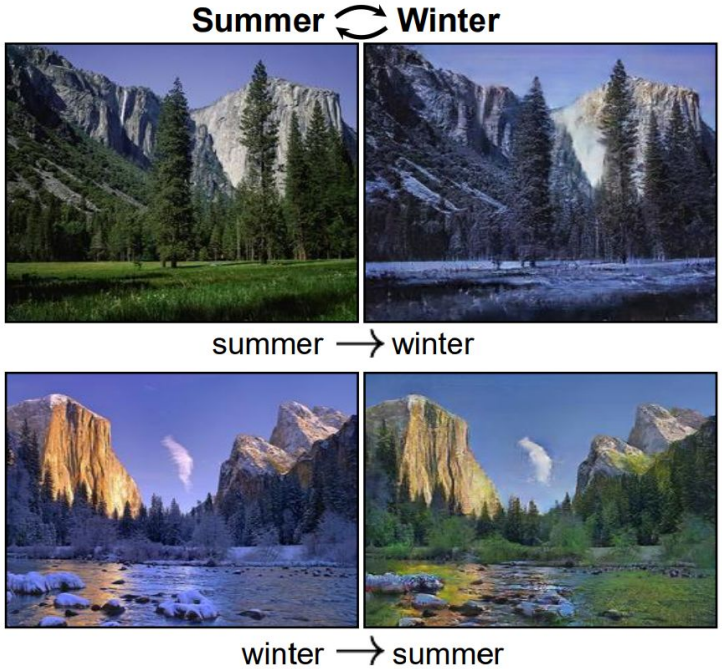
Jun-Yan Zhu* **Taesung Park*** **Phillip Isola** **Alexei A. Efros**

UC Berkeley

In ICCV 2017

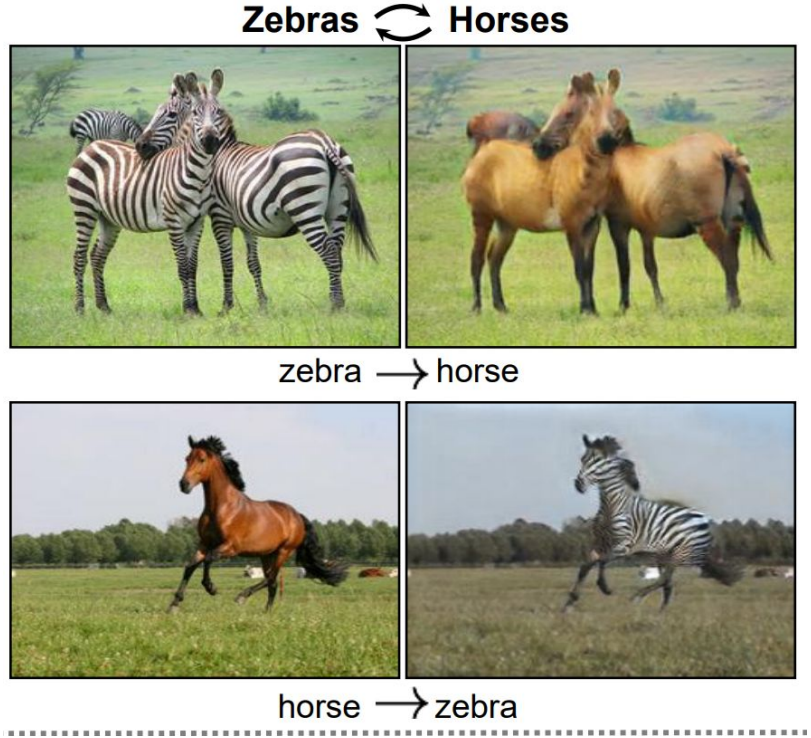
Motivación

- Los ejemplos en pares podrían ser:
  - **Difíciles de conseguir**



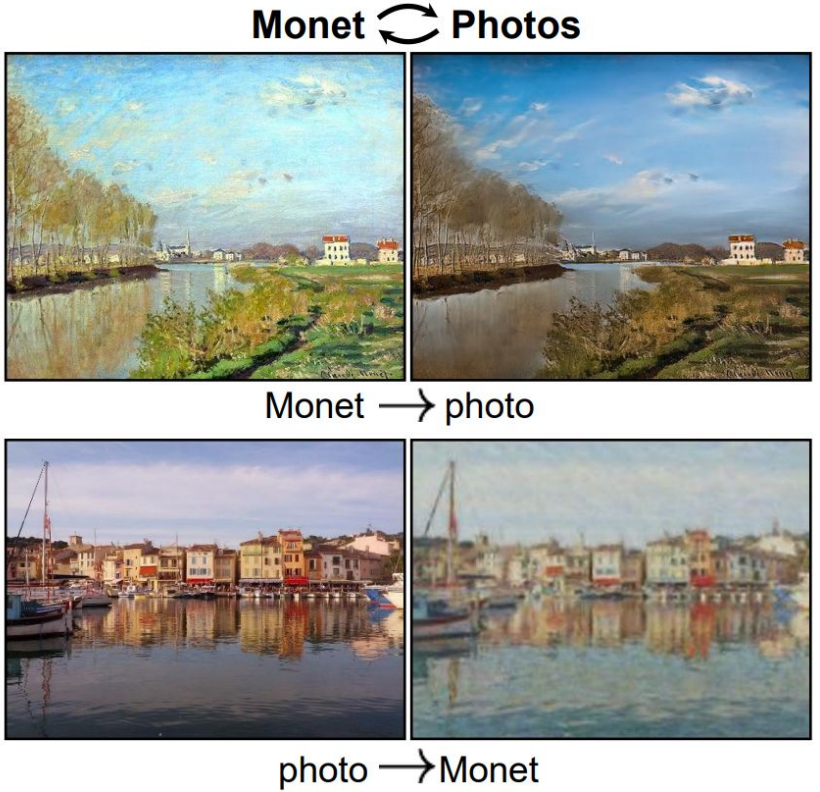
Motivación

- Los ejemplos en pares podrían ser:
  - Difíciles de conseguir
  - **Muy difíciles de conseguir**



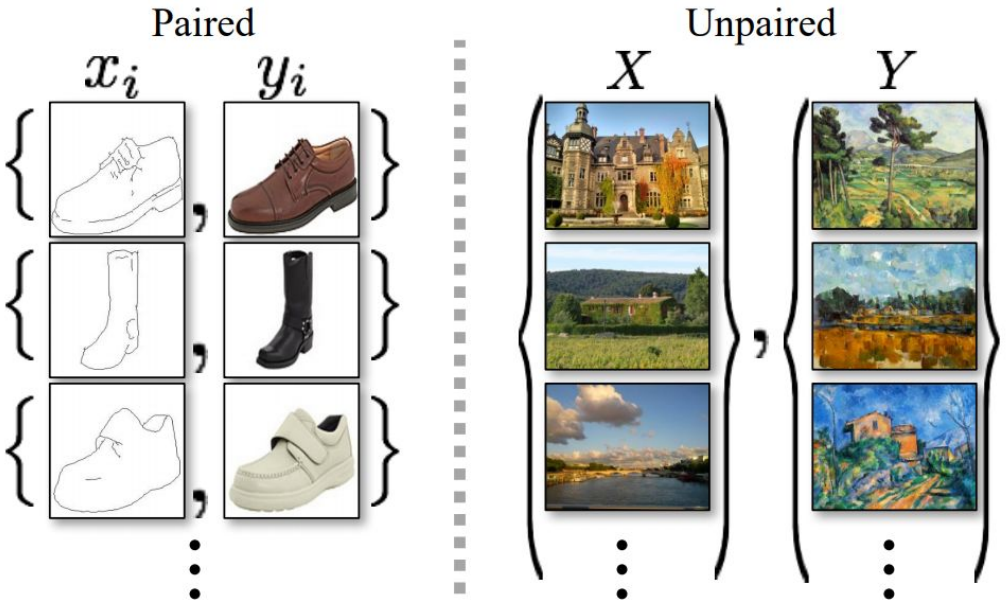
Motivación

- Los ejemplos en pares podrían ser:
  - Difíciles de conseguir
  - Muy difíciles de conseguir
  - **¡Ni siquiera existir!**



Solución: usar la información de los conjuntos

Imágenes no pareadas pero usándolas como conjuntos, pueden proveer una supervisión débil

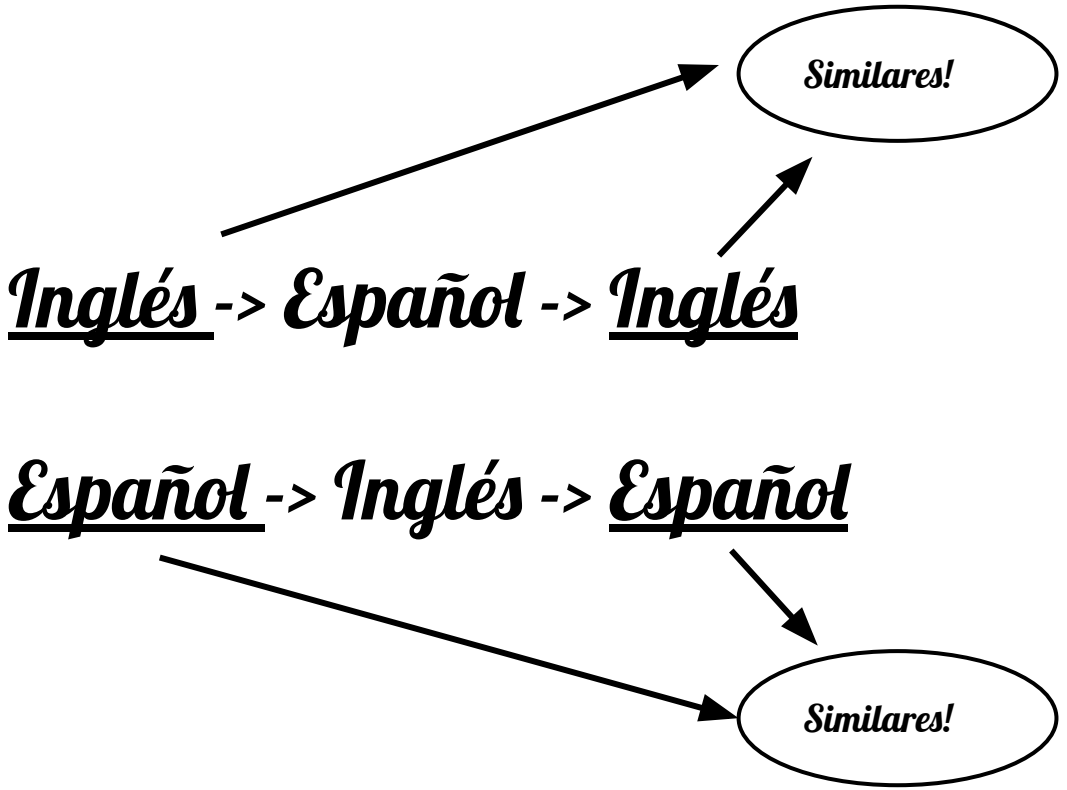


La idea de la consistencia de ciclos

Inglés -> Español -> Inglés

Español -> Inglés -> Español

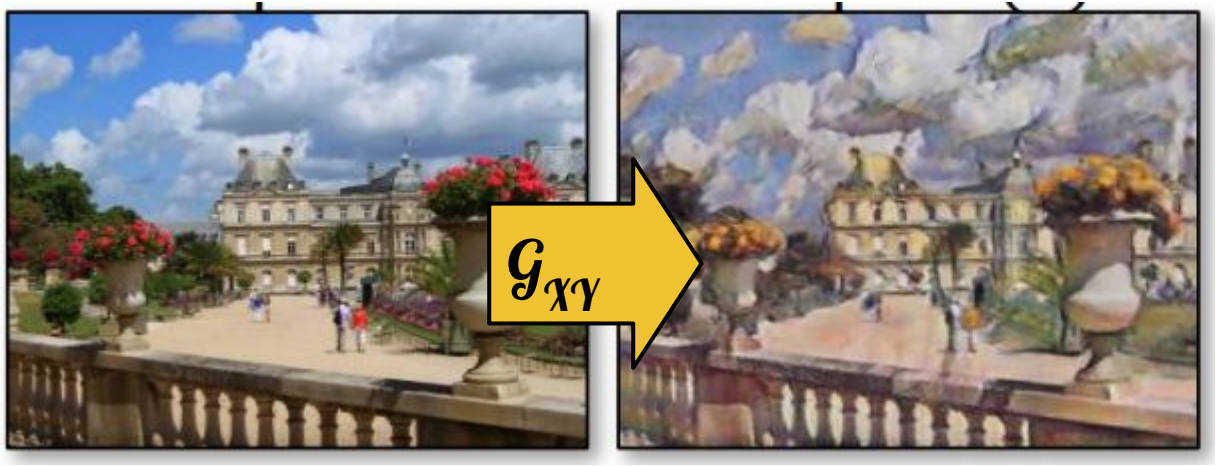
La idea de la consistencia de ciclos



Ahora con imágenes reales

x

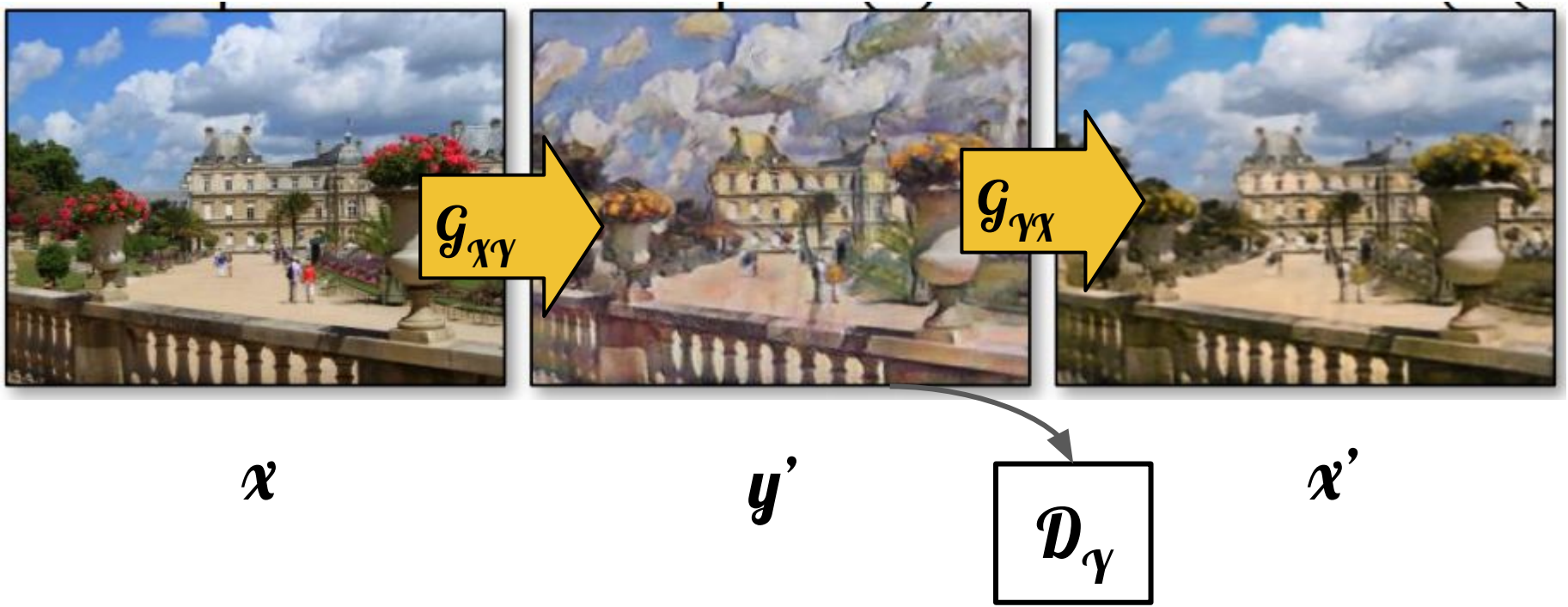
Ahora con imágenes reales



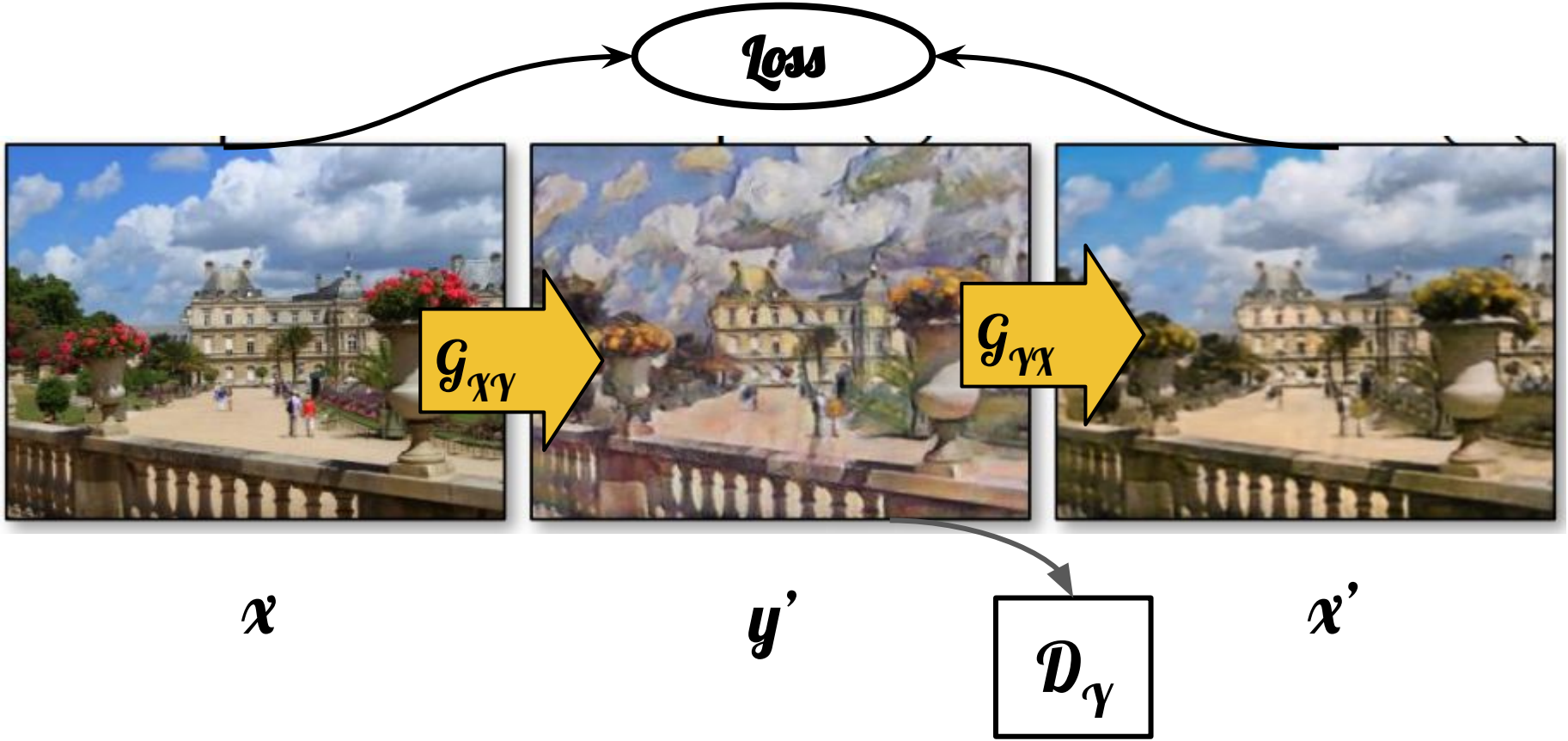
x y'

Ahora con imágenes reales

Ahora con imágenes reales

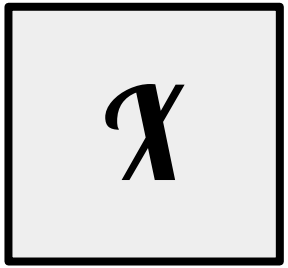


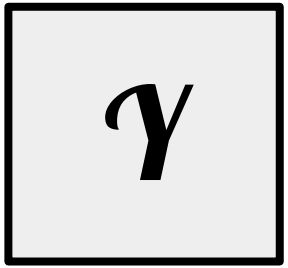
Ahora con imágenes reales



Pensemos en dos mapeos (o generadores)

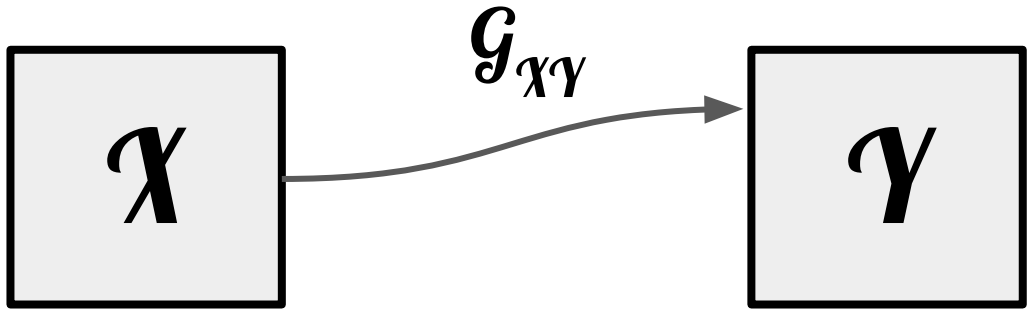
Consideremos dos dominios de imágenes X e Y





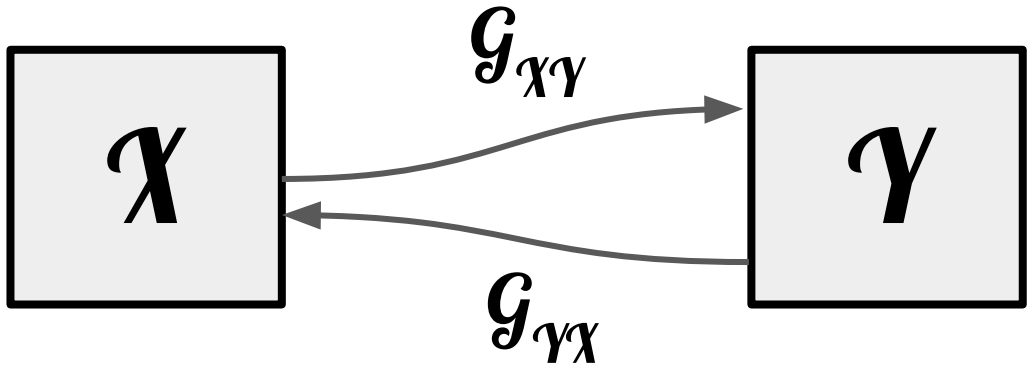
Pensemos en dos mapeos (o generadores)

Podemos tener un mapeo de X a Y



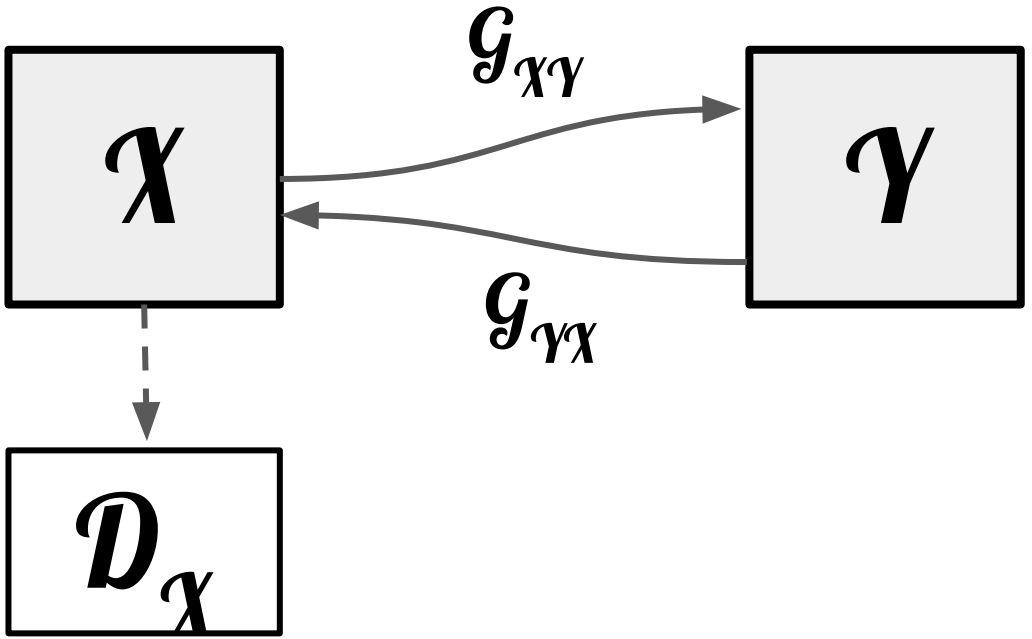
Pensemos en dos mapeos (o generadores)

A la vez, podemos tener un mapeo de Y a X



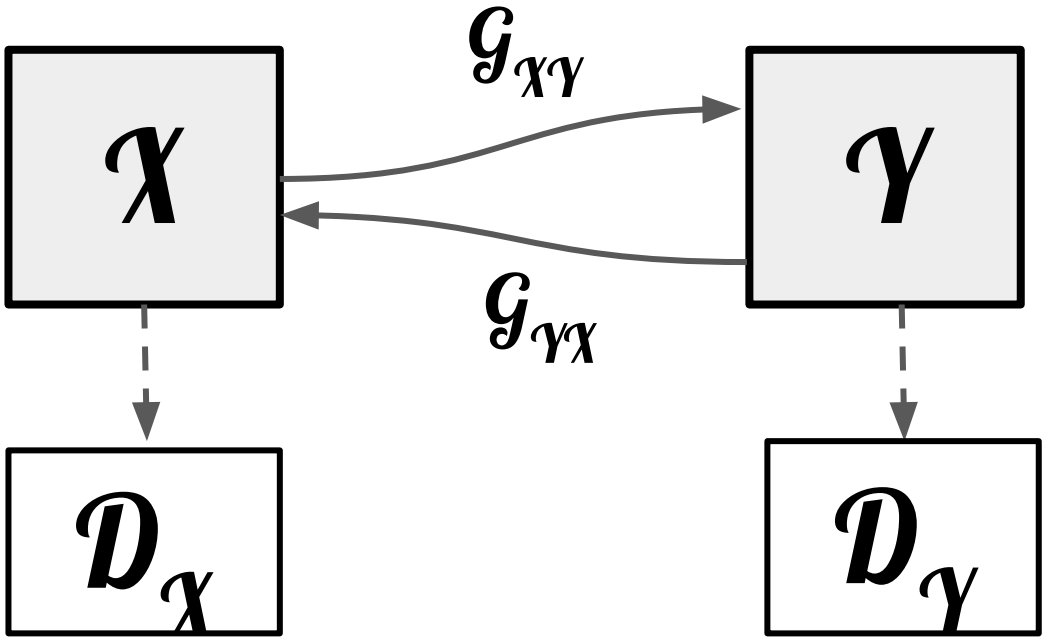
...también en dos discriminadores

Un discriminador para el dominio X

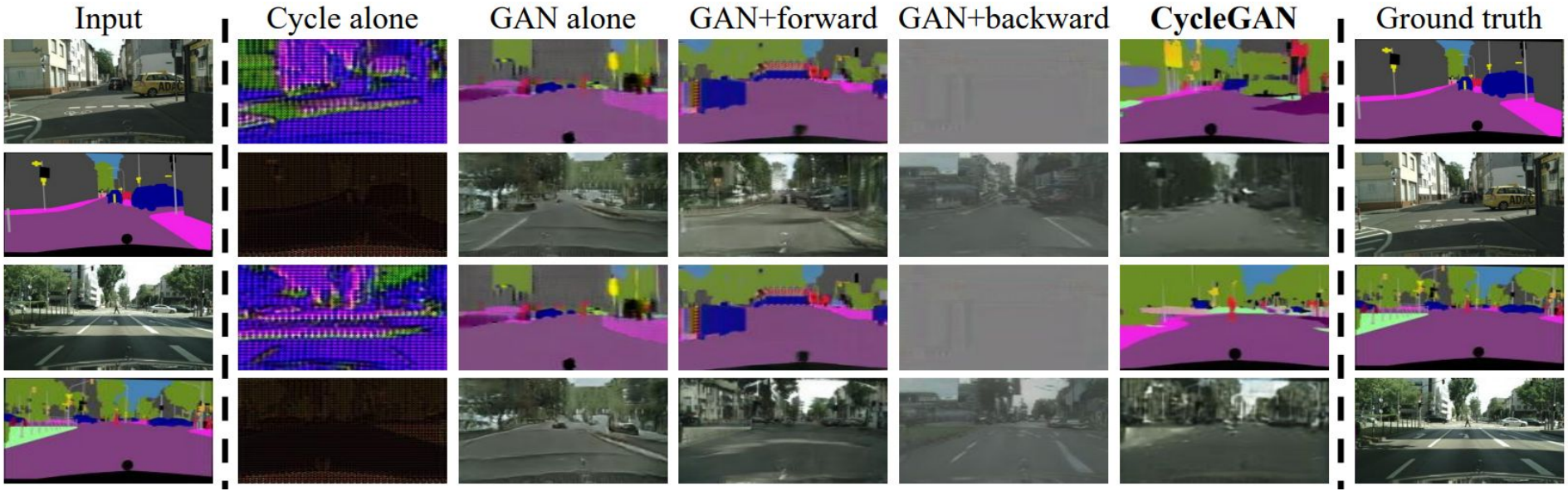


...también en dos discriminadores

Un discriminador para el dominio Y



Todos los componentes... visualmente



The full objective is:

$$\mathcal{L}(G_{XY}, G_{YX}, D_X, D_Y) = \mathcal{L}_{\text{GAN}}(G_{XY}, X, Y, D_Y) +$$

Going from X to Y produces something in the same distribution as Y

The full objective is:

$$\mathcal{L}(G_{XY}, G_{YX}, D_X, D_Y) = \mathcal{L}_{\text{GAN}}(G_{XY}, X, Y, D_Y) +$$

$$\boxed{\mathcal{L}_{\text{GAN}}(G_{YX}, Y, X, D_X) +}$$

Going from Y to X produces something in the same distribution as X

The full objective is:

$$\mathcal{L}(G_{XY}, G_{YX}, D_X, D_Y) = \mathcal{L}_{\text{GAN}}(G_{XY}, X, Y, D_Y) +$$

$$\mathcal{L}_{\text{GAN}}(G_{YX}, Y, X, D_X) +$$

$$\lambda \mathcal{L}_{\text{cyc}}(G_{XY}, G_{YX})$$

A reconstruction loss of the cycle, weighted by λ, for both directions. This can be seen as filling the role of the structural loss in pix2pix

The full objective is:

$$\mathcal{L}(G_{XY}, G_{YX}, D_X, D_Y) = \mathcal{L}_{\text{GAN}}(G_{XY}, X, Y, D_Y) +$$

$$\mathcal{L}_{\text{GAN}}(G_{YX}, Y, X, D_X) +$$

$$\lambda \mathcal{L}_{\text{cyc}}(G_{XY}, G_{YX})$$

Usually λ has a value of 10, while in pix2pix the value is 100

Traducción de Imágenes: Pix2PixHD

07

High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs

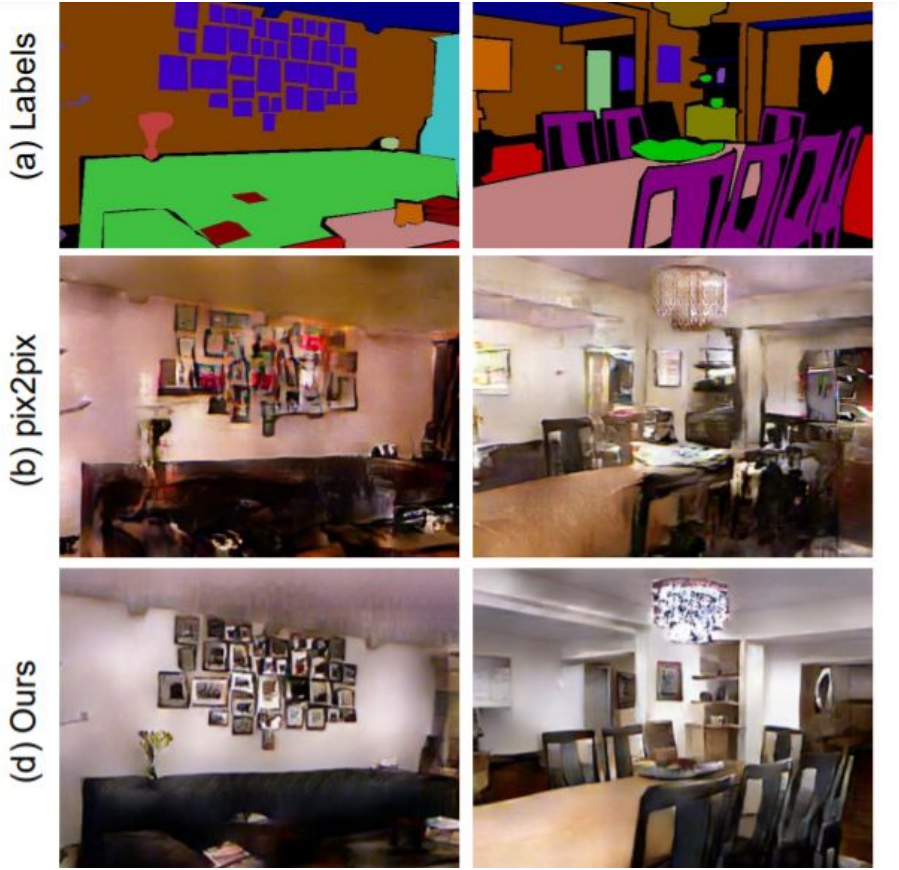
Ting-Chun Wang[1] Ming-Yu Liu[1] Jun-Yan Zhu[2] Andrew Tao[1] Jan Kautz[1] Bryan Catanzaro[1]
[1]NVIDIA Corporation [2]UC Berkeley

Overview

- El problema de generación de imágenes condicionada con una imagen de entrada, puede usarse para síntesis de imágenes a partir de las etiquetas

Overview

- Es la continuación de Pix2Pix, generando imágenes de mayor resolución



VGG/Perceptual Loss

- Una forma de generar imágenes con alta resolución es usar el espacio inducido por una red VGG entrenada para ImageNet y calcular la distancia en este espacio

$$\| \text{VGG}(Y) - \text{VGG}(G(X)) \|_1$$

- Se entiende que este espacio es más significativo “perceptualmente”

Discriminador Multiescala

- Para generar alta resolución se necesita evaluar distintos tamaños de campo receptivo
- Se usan 3 discriminadores (D1, D2 y D3) que tienen una estructura idéntica pero operan en imágenes de diferente tamaño

$$\min_{G} \max_{D_1, D_2, D_3} \sum_{k=1,2,3} \mathcal{L}_{\text{GAN}}(G, D_k)$$

¿Qué genera la alta resolución?

- Es la continuación de Pix2Pix
- Los ejemplos de entrenamiento son pareados (igual a Pix2Pix, a diferencia de CycleGAN)

Pablo Fonseca
pfonseca@pucp.edu.pe
https://sites.google.com/view/palefo