Conditional GAN

Deep Image Processing Seminar

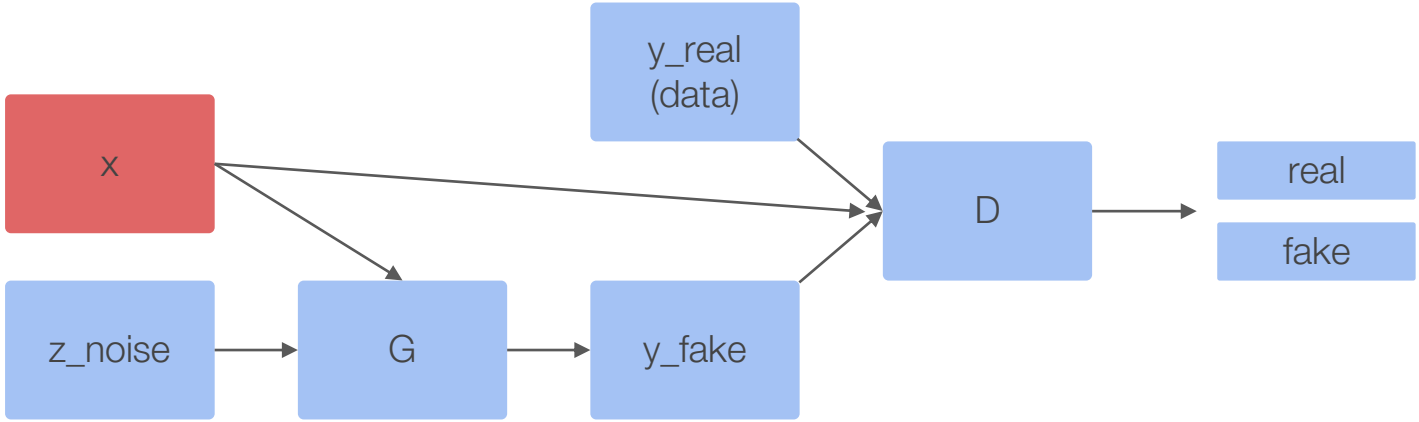
Agenda

- Conditional Gan (Nov 2014)
- Image-to-Image Translation (Nov 2017)
- Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks (Feb 2018)

Conditional Gan

Mehdi Mirza,
Simon Osindero
Nov 2014

Conditional GAN



Conditional Gan, Mehdi Mirza and Simon Osindero (Nov 2014)

Why do we need a conditional GAN?

In an unconditioned generative model, there is no control on modes of the data being generated.

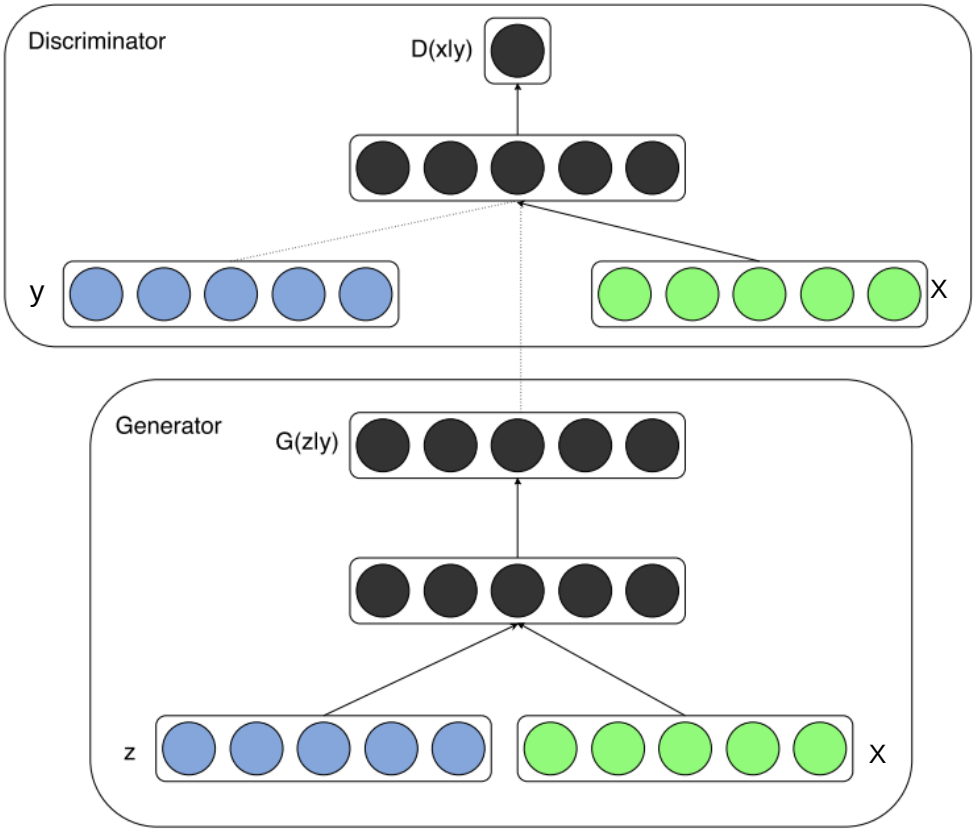
In the Conditional GAN (CGAN), the generator learns to generate a fake sample with a specific condition or characteristics (such as a label associated with an image or more detailed tag) rather than a generic sample from unknown noise distribution.

Conditional generative adversarial nets for convolutional face generation
Jon Gauthier

Conditional GAN

In the discriminator x and y are presented as inputs to a discriminative function.

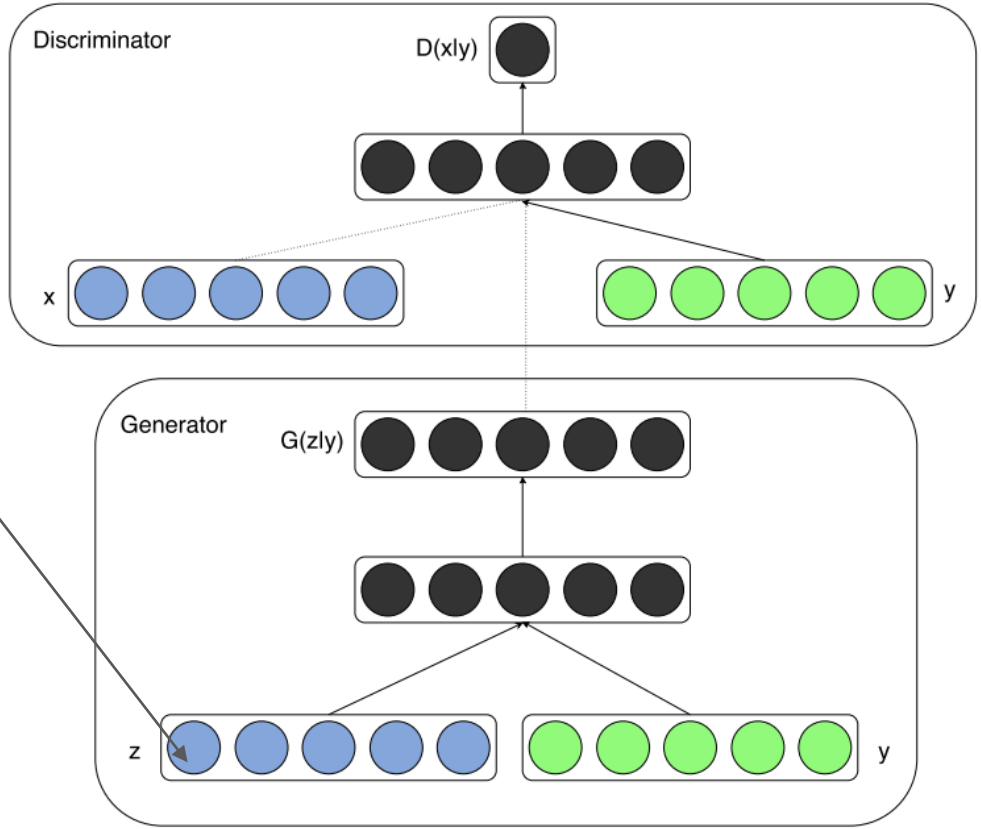
In the generator the prior input noise z, and y are combined in joint hidden representation.



Conditional Gan, Mehdi Mirza and Simon Osindero (Nov 2014)

Conditional GAN

Why do we still need z?



GAN min-max reminder

$$\min_G \max_D V(G,D) = \mathbb{E}_{y \sim p_{data}(y)}[logD(y)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))]$$

Train discriminator to maximize the probability of the training data

Train the discriminator to minimize the probability of the data sampled from the generator

Train the generator to maximize the probability that the discriminator assigns to its own sample

Conditional GAN min-max

$$\min_G \max_D \left(\mathbb{E}_{y,x\sim p_{data}(y,x)}[logD(y,x)] + \mathbb{E}_{x\sim p_x, z\sim p_{z(z)}}[\log(1 - D(G(z,x),x))]\right)$$

Train discriminator to maximize the probability of the training data

Train the discriminator to minimize the probability of the data sampled from the generator

Train the generator to maximize the probability that the discriminator assigns to its own sample

Conditional Gan, Mehdi Mirza and Simon Osindero (Nov 2014)

Experimental Results

$[1,0,0,0,0,0,0,0,0,0] \rightarrow$

$[0,1,0,0,0,0,0,0,0,0] \rightarrow$

$[0,0,1,0,0,0,0,0,0,0] \rightarrow$

$[0,0,0,1,0,0,0,0,0,0] \rightarrow$

$[0,0,0,0,1,0,0,0,0,0] \rightarrow$

$[0,0,0,0,0,1,0,0,0,0] \rightarrow$

$[0,0,0,0,0,0,1,0,0,0] \rightarrow$

$[0,0,0,0,0,0,0,1,0,0] \rightarrow$

$[0,0,0,0,0,0,0,0,1,0] \rightarrow$

$[0,0,0,0,0,0,0,0,0,1] \rightarrow$

Experimental Results - auto tagging of images

	User tags + annotations	Generated tags
	montanha, trem, inverno, frio, people, male, plant life, tree, structures, transport, car	taxi, passenger, line, transportation, railway station, passengers, railways, signals, rail, rails
	food, raspberry, delicious, homemade	chicken, fattening, cooked, peanut, cream, cookie, house made, bread, biscuit, bakes
	water, river	creek, lake, along, near, river, rocky, treeline, valley, woods, waters
	people, portrait, female, baby, indoor	love, people, posing, girl, young, strangers, pretty, women, happy, life

Table 2: Samples of generated tags

UCM
User generated metadata

Image-to-Image Translation

Philip Isola, Jun-Yan Zhu
Tinghui Zhou, Alexei A. Efros
Nov 2017

Let’s start with a demo

Image to image mapping

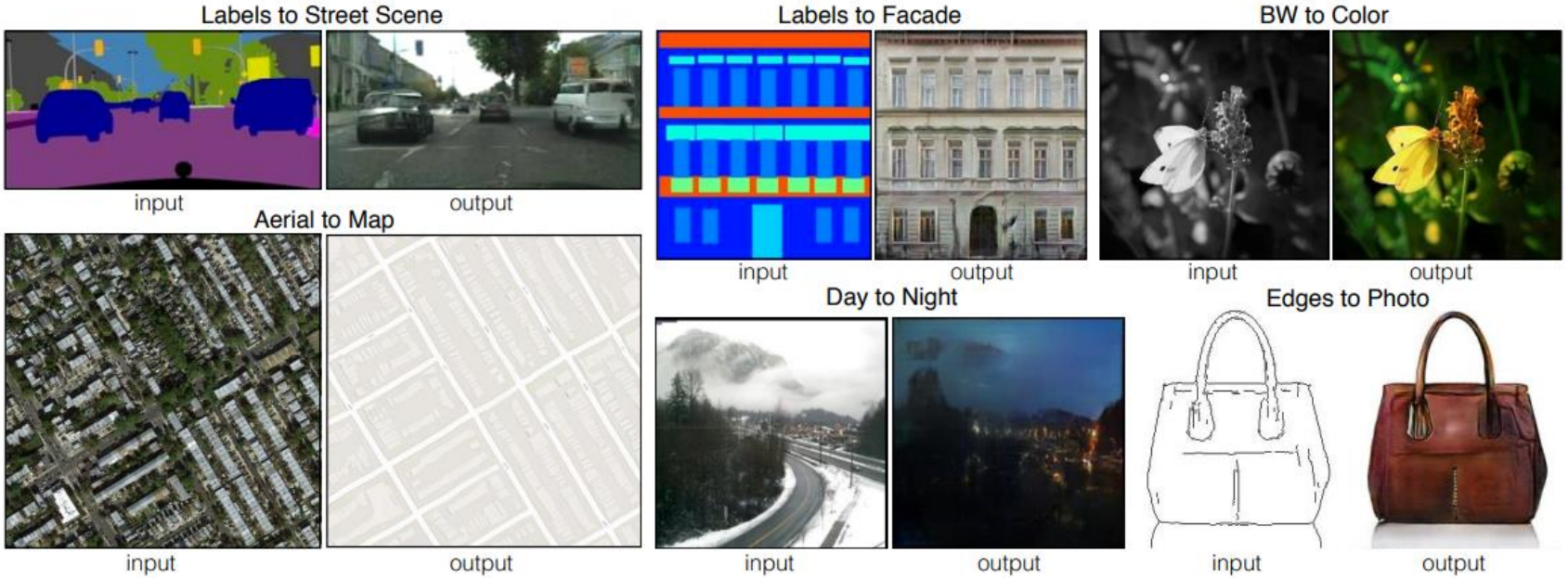


Image-to-Image Translation, Philip Isola, Jun-Yan Zhu, Tinghui Zhou, Alexei A. Efros (Nov 2017)

Generalization

- Setting is always the same: predicting pixels from pixels
- Conditional adversarial nets are a general-purpose solution that appears to work well on a wide variety of these problems. Same architecture and objective, and simply train on different data.

- Similar direction: CNNs.
  Problem: designing effective losses.
  Loss: “Minimize Euclidean distance between predicted and ground truth”
- Better:
  Goal: “make output indistinguishable from reality”
  Loss: learn automatically from goal (GAN!)
- GANs loss functions adapt to data

Image → image transformations.
Why do we need a **conditional** GAN for that?

Conditional generative model:
We'll condition on the input image

Method - training process

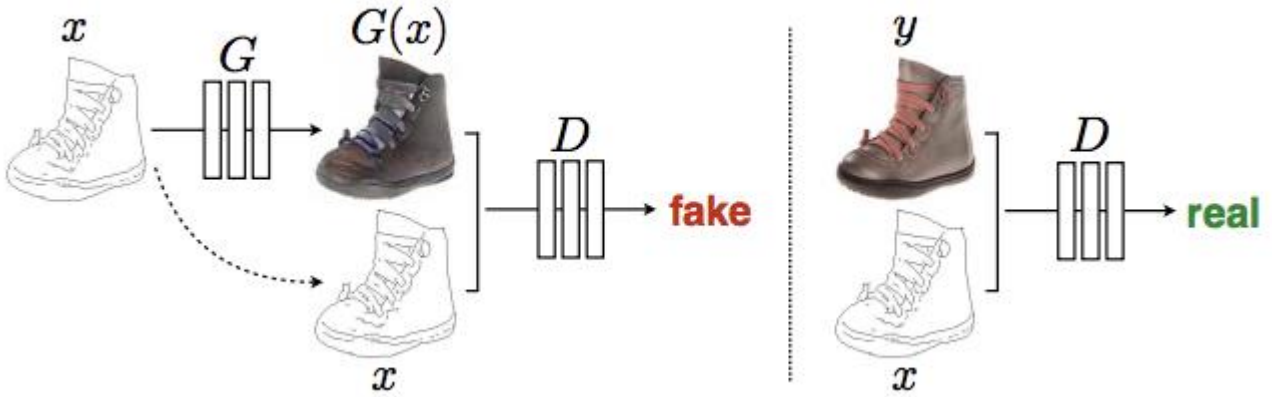


Image-to-Image Translation, Philip Isola, Jun-Yan Zhu, Tinghui Zhou, Alexei A. Efros (Nov 2017)

Objective

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

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

Tested importance of conditioning the discriminator:

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

Mixing the GAN objective

Previous approaches found it beneficial to mix the GAN objective with a traditional loss function.

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

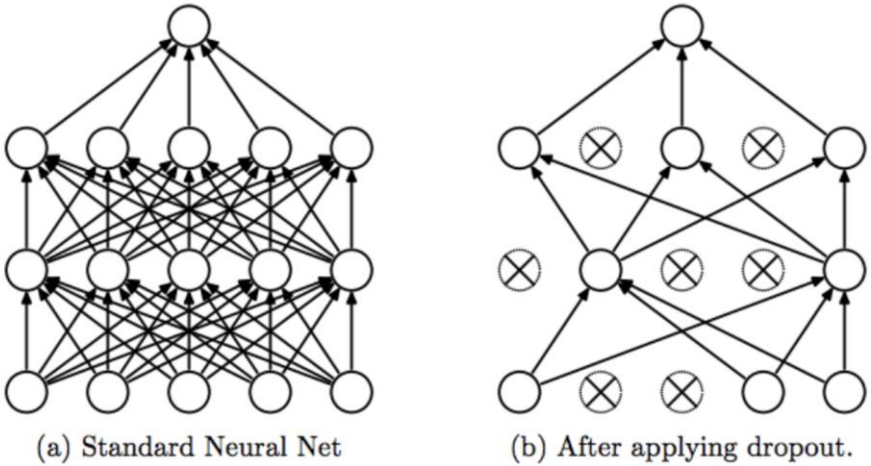
$$G^* = \arg\min_G \max_D \mathcal{L}_{cGAN}(G,D) + \lambda \mathcal{L}_{L1}(G)$$

Dropout

Addressing overfitting concerns:

- Using z, past cGANs used Gaussian noise as z
- Here: provide noise only in form of dropout

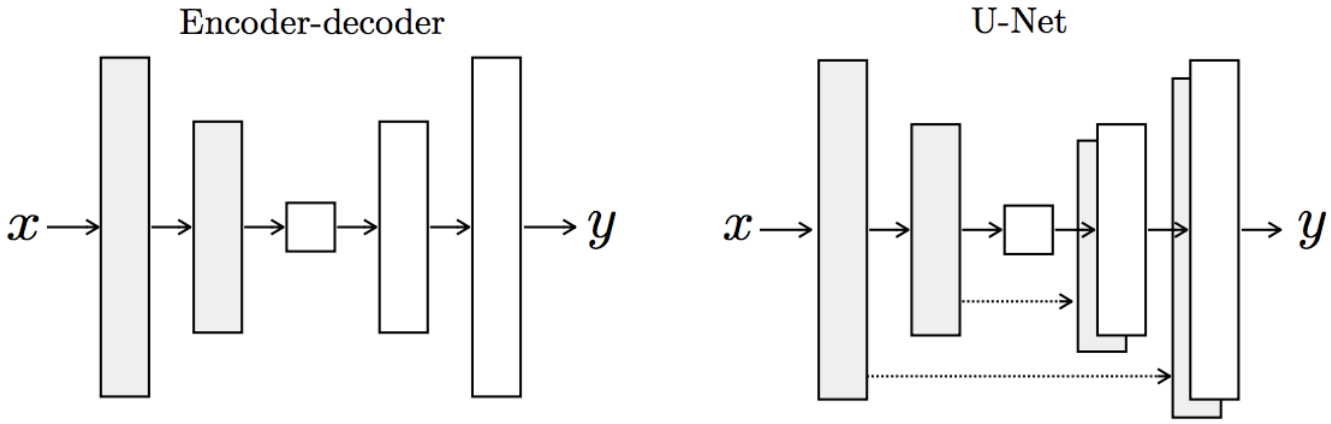
Still:
Output not stochastic enough



(a) Standard Neural Net (b) After applying dropout.

Srivastava, Nitish, et al. "Dropout: a simple way to prevent neural networks from overfitting", JMLR 2014

Network Architectures - Generator



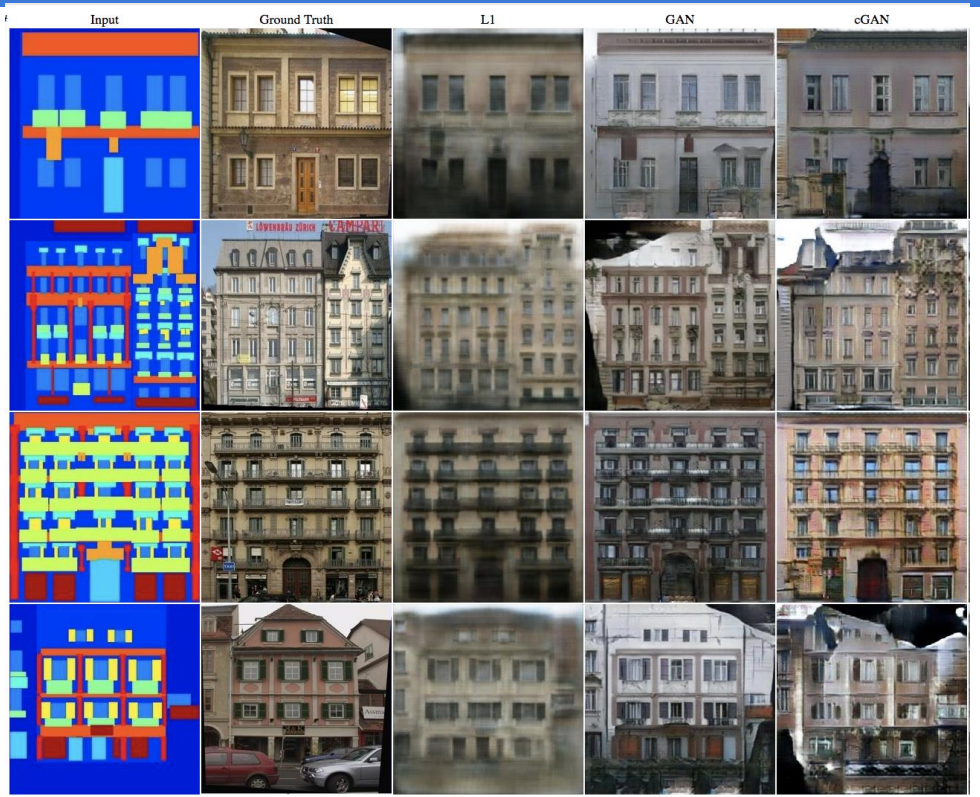
Network Architectures - Discriminator

For low frequencies, L1/L2 loss will suffice for enforcing correctness.

We'll restrict the GAN discriminator to only model high frequencies.

↓

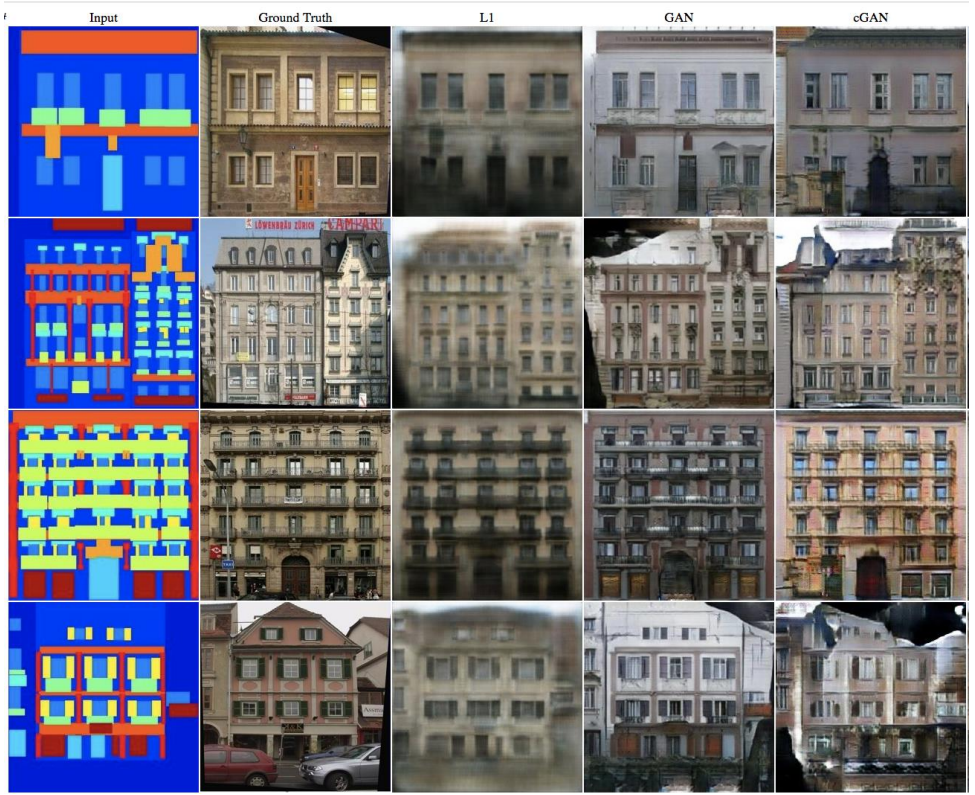
Sufficient to restrict attention only to local image patches:
PatchGAN



Network Architectures - Discriminator

PatchGAN

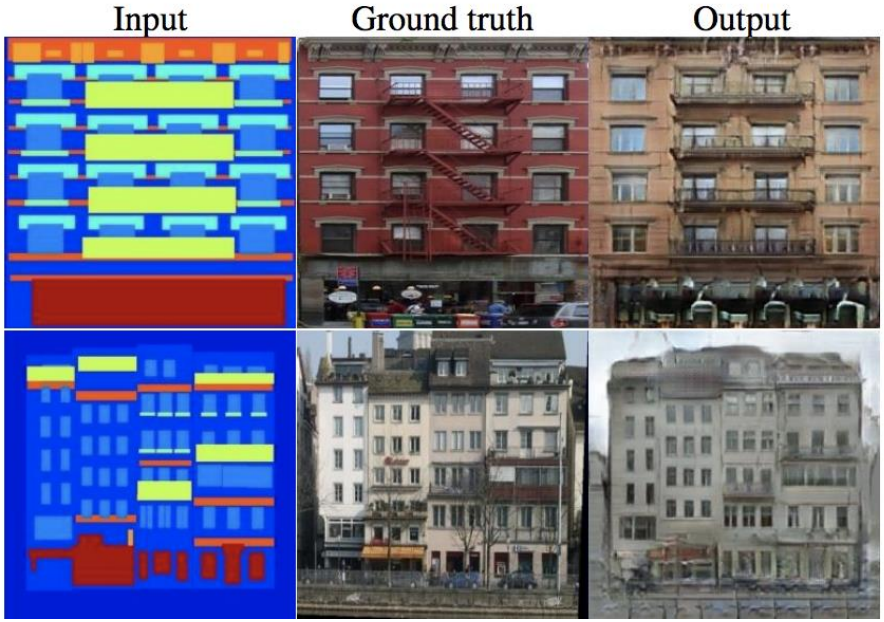
- Classifies if each N*N patch in the image is real or fake
- This discriminator is run convolutionally across the image, averaging all responses to provide D's output.
- Patch is smaller than image



Experiments and Evaluation

Experiments and Evaluation

- *Semantic labels↔photo*, trained on the Cityscapes dataset [11].
- *Architectural labels→photo*, trained on CMP Facades [44].
- *Map↔aerial photo*, trained on data scraped from Google Maps.
- *BW→color photos*, trained on [50].
- *Edges→photo*, trained on data from [64] and [59]; binary edges generated using the HED edge detector [57] plus postprocessing.
- *Sketch→photo*: tests edges→photo models on human-drawn sketches from [18].
- *Day→night*, trained on [32].
- *Thermal→color photos*, trained on data from [26].
- *Photo with missing pixels→inpainted photo*, trained on Paris StreetView from [13].



400 images in dataset, less than 2 hours of training

Experiments and Evaluation

Experiments and Evaluation

Experiments and Evaluation

Experiments and Analysis

How do you evaluate synthesized images?

- Plausibility to a human observed is the ultimate goal →
  With actual people: using AMT
- Measuring to see if off-the-shelf recognition systems can recognize them →
  FCN-score, using pre-training semantic classifiers

	Photo → Map	**Map → Photo**
Loss	% Turkers labeled *real*	% Turkers labeled *real*
L1	2.8% ± 1.0%	0.8% ± 0.3%
L1+cGAN	6.1% ± 1.3%	**18.9% ± 2.5%**

Table 4: AMT "real vs fake" test on maps↔aerial photos.

Method	% Turkers labeled *real*
L2 regression from [61]	16.3% ± 2.4%
Zhang et al. 2016 [61]	**27.8% ± 2.7%**
Ours	22.5% ± 1.6%

Table 5: AMT "real vs fake" test on colorization.

Loss	**Per-pixel acc.**	**Per-class acc.**	**Class IOU**
Encoder-decoder (L1)	0.35	0.12	0.08
Encoder-decoder (L1+cGAN)	0.29	0.09	0.05
U-net (L1)	0.48	0.18	0.13
U-net (L1+cGAN)	**0.55**	**0.20**	**0.14**

Objective function analysis

- L1 alone - blurry, greyish
- cGAN alone - imaginative
  Better with colors
- Removing conditioning data from the discriminator resulted in: only cares that the output is realistic

Image-to-Image Translation, Philip Isola, Jun-Yan Zhu, Tinghui Zhou, Alexei A. Efros (Nov 2017)

Generator analysis

U-Net architecture allowed low-level information to shortcut across the network.
Did this improve results?

Loss	Per-pixel acc.	Per-class acc.	Class IOU
Encoder-decoder (L1)	0.35	0.12	0.08
Encoder-decoder (L1+cGAN)	0.29	0.09	0.05
U-net (L1)	0.48	0.18	0.13
U-net (L1+cGAN)	**0.55**	**0.20**	**0.14**

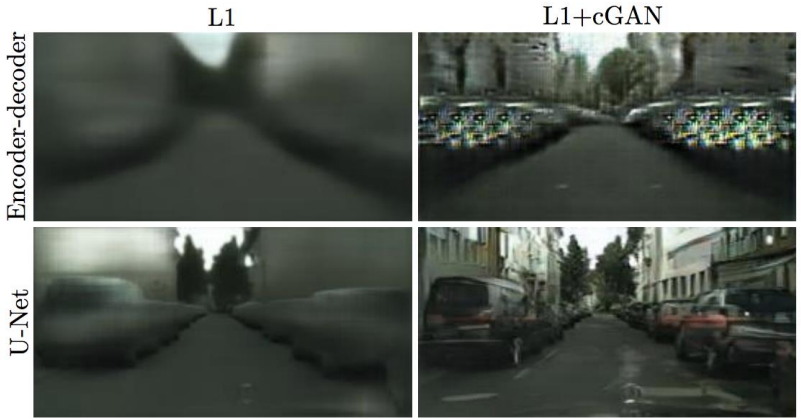


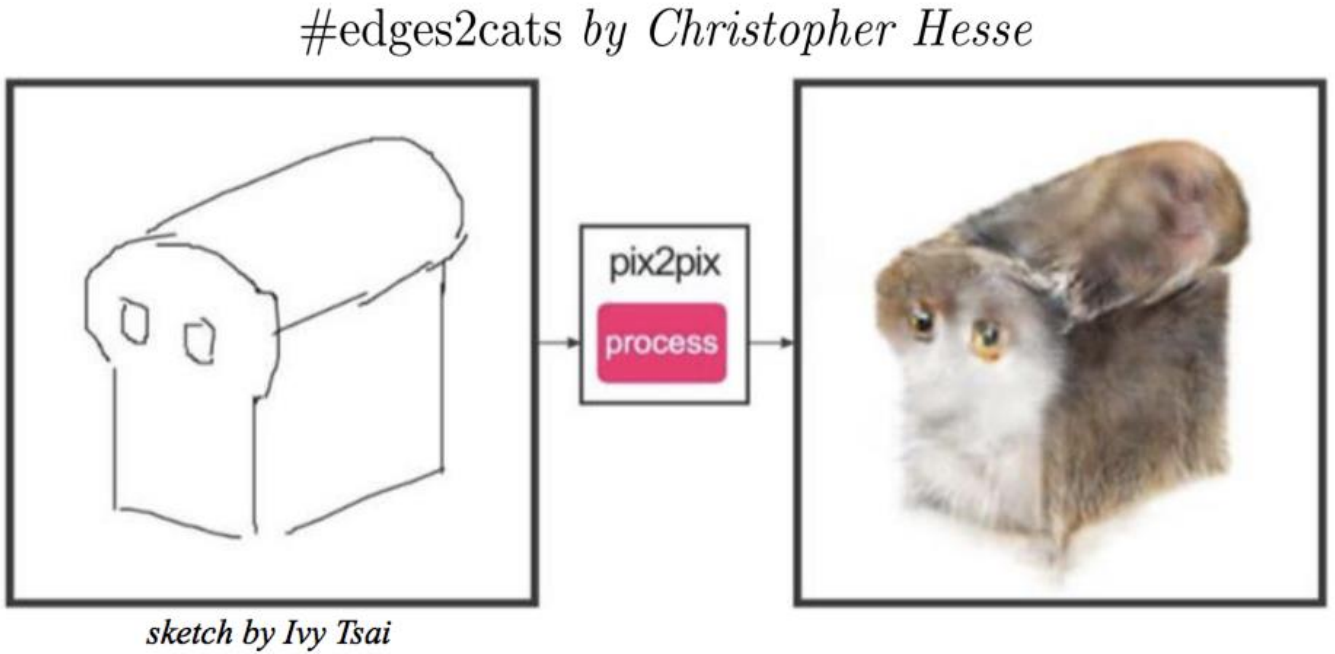
Image-to-Image Translation, Philip Isola, Jun-Yan Zhu, Tinghui Zhou, Alexei A. Efros (Nov 2017)

Discriminator analysis

PatchGAN
Tested the effect of varying the patch size.

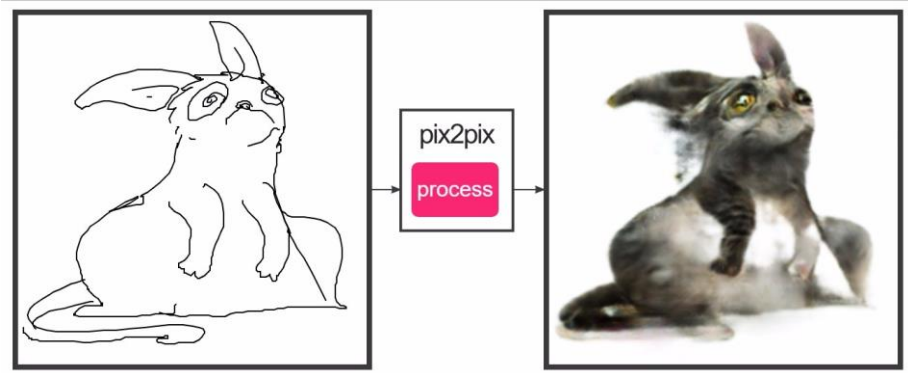
Discriminator receptive field	Per-pixel acc.	Per-class acc.	Class IOU
1×1	0.39	0.15	0.10
16×16	0.65	0.21	**0.17**
70×70	**0.66**	**0.23**	**0.17**
286×286	0.42	0.16	0.11

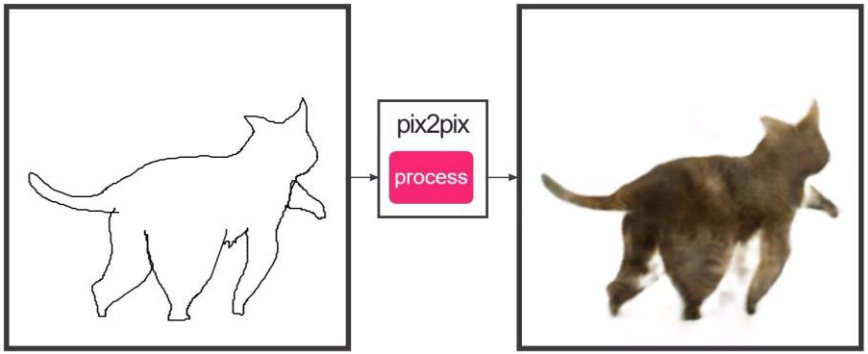
Organic Applications



Twitter #edges2cats

Organic Applications





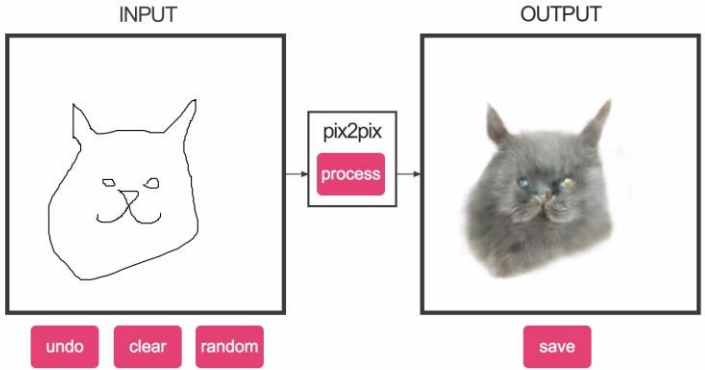


Image-to-Image Translation, Philip Isola, Jun-Yan Zhu, Tinghui Zhou, Alexei A. Efros (Nov 2017)

Organic Applications

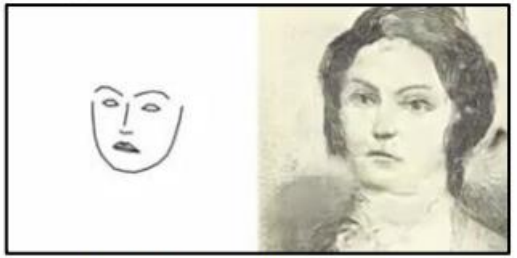
Background removal

by Kaihu Chen

Palette generation

by Jack Qiao

Sketch→ Portrait



by Mario Klingemann

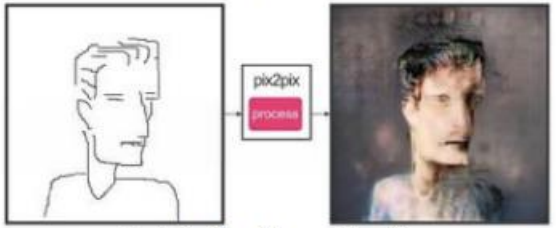
Sketch → Pokemon

by Bertrand Gondouin

"Do as I do"

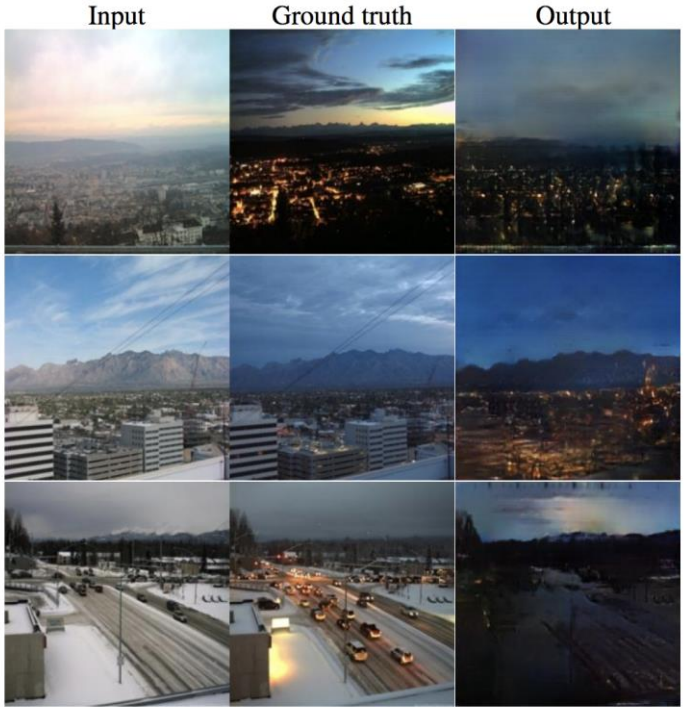
by Brannon Dorsey

#fotogenerator



sketch by Yann LeCun

More results



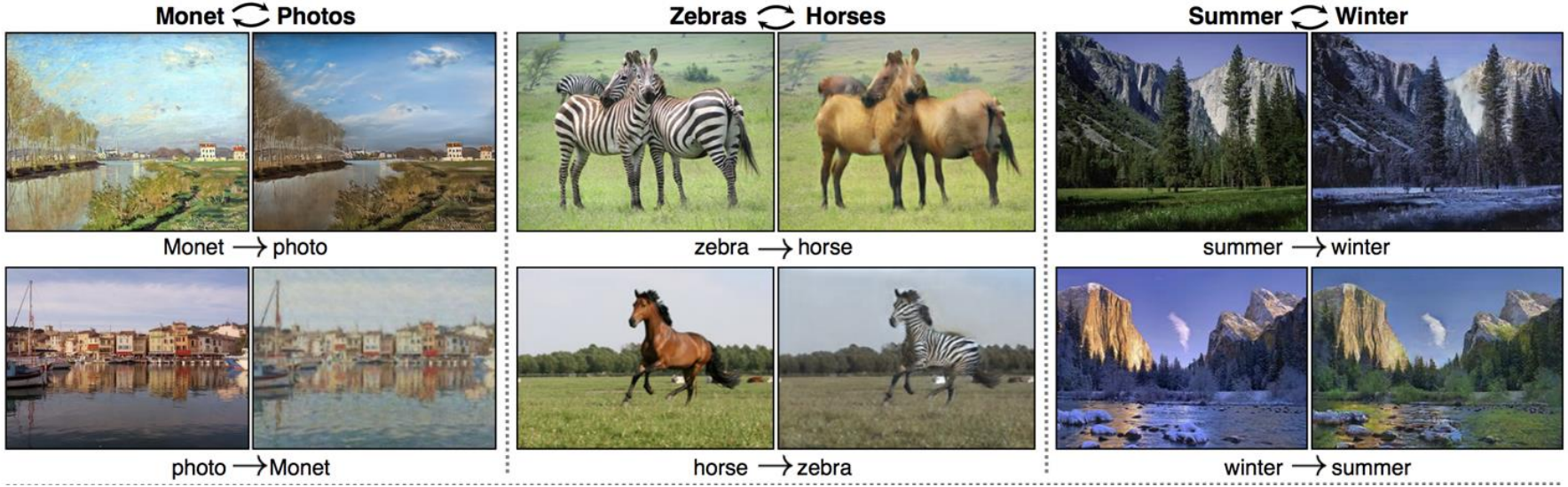
5. Conclusion

The results in this paper suggest that conditional adversarial networks are a promising approach for many image-to-image translation tasks, especially those involving highly structured graphical outputs. These networks learn a loss adapted to the task and data at hand, which makes them applicable in a wide variety of settings.

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

Jun-Yan Zhu, Taesung Park, Phillip Isola, Alexei A. Efros
Feb 2018

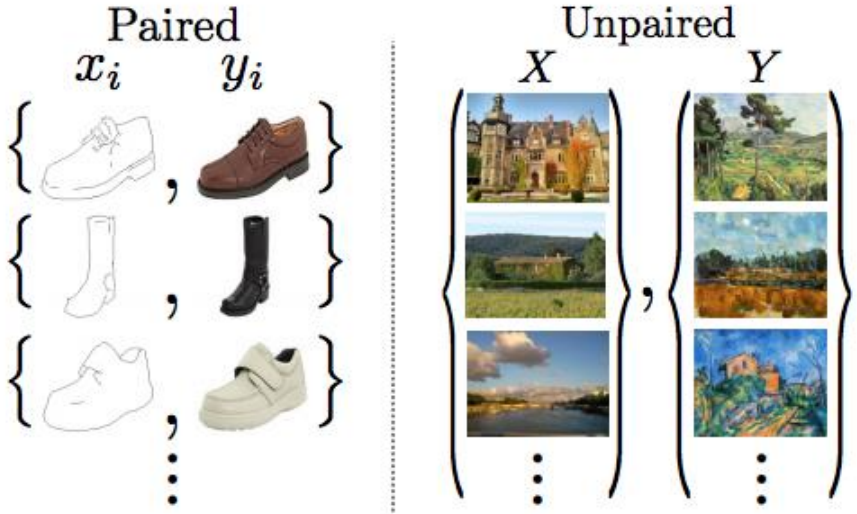
Unpaired Image to Image



Given any two unordered image collections X and Y, our algorithm learns to automatically “translate” an image from one into the other and vice versa

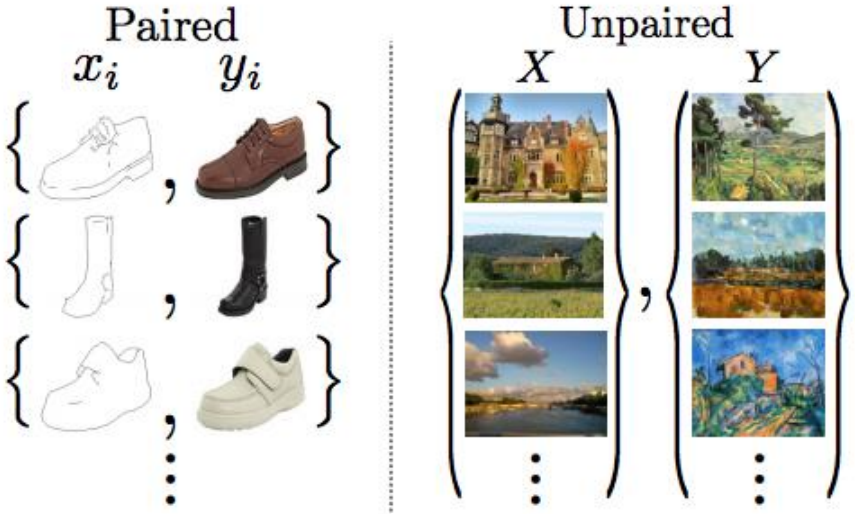
Unpaired image-to-Image Translation, Jun-Yan Zhu, Taesung Park, Phillip Isola, Alexei A. Efros (Feb 2018)

Unpaired Image to Image



We are **not** using a training set of aligned image pairs
We'll translate an image from a source domain X to a target domain Y in the absence of paired examples

Unpaired Image to Image



“Our approach builds on the “pix2pix” framework of Isola et al., which uses a conditional generative adversarial network to learn a mapping from input to output images”

Unpaired Image to Image

We want:
Mapping $G : X \rightarrow Y$ such that the output $\hat{y} = G(x), x \in X$ is indistinguishable from images $y \in Y$ by an adversary trained to classify $\hat{y}$ apart from y.

Problems:
Meaningful pairing
In practice - mode collapse

We need to add more structure to our objective

Unpaired Image to Image

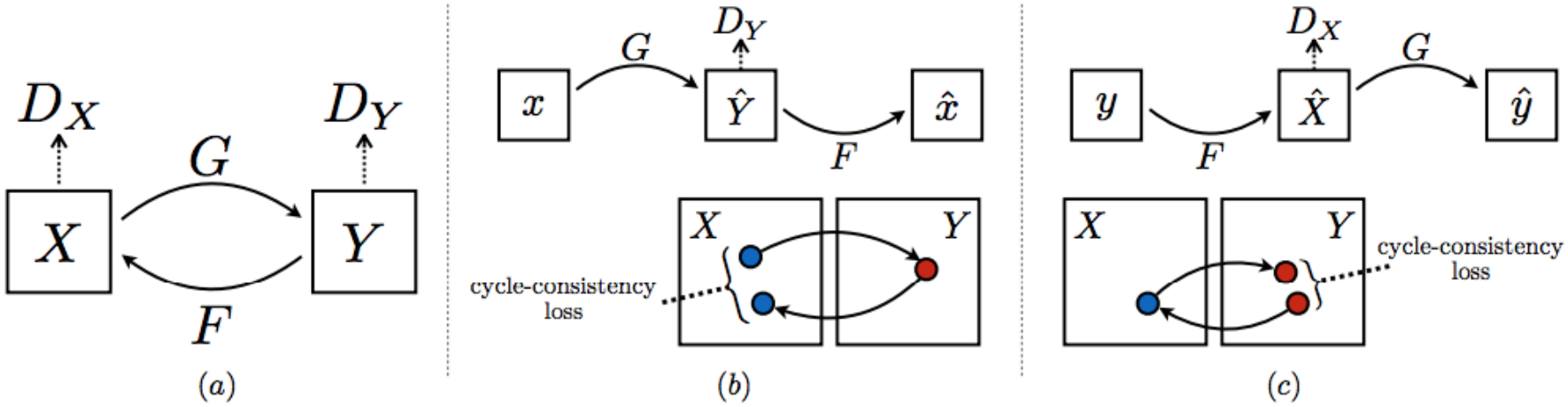
We need to add more structure to our objective →
Translation should be “cycle consistent”

$$F : Y \rightarrow X \quad G : X \rightarrow Y$$

$$F(G(x)) \approx x \text{ and } G(F(y)) \approx y$$

Combining this loss with adversarial losses on domains X and Y yields our full objective for unpaired image-to-image translation

Consistency Cycle loss



Unpaired image-to-Image Translation, Jun-Yan Zhu, Taesung Park, Phillip Isola, Alexei A. Efros (Feb 2018)

Consistency Cycle loss

Adversarial losses:

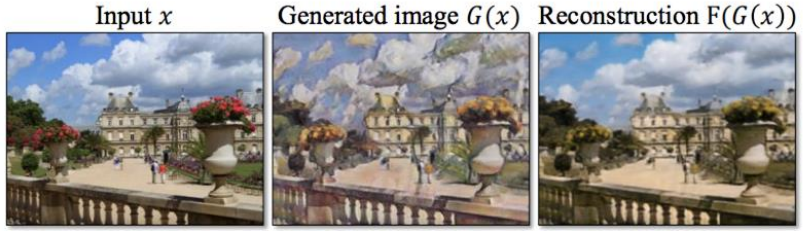
$$G : X \rightarrow Y$$

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

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

Cycle consistency loss:

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



Full objective

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

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

Experiments and Evaluation

Experiments and Analysis

How do you evaluate synthesized images?

- Plausibility to a human observed is the ultimate goal →
  With actual people: using AMT
- Measuring to see if off-the-shelf recognition systems can recognize them →
  FCN-score, using pre-training semantic classifiers

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

Experiments and Analysis

Unpaired image-to-Image Translation, Jun-Yan Zhu, Taesung Park, Phillip Isola, Alexei A. Efros (Feb 2018)

Experiments and Analysis

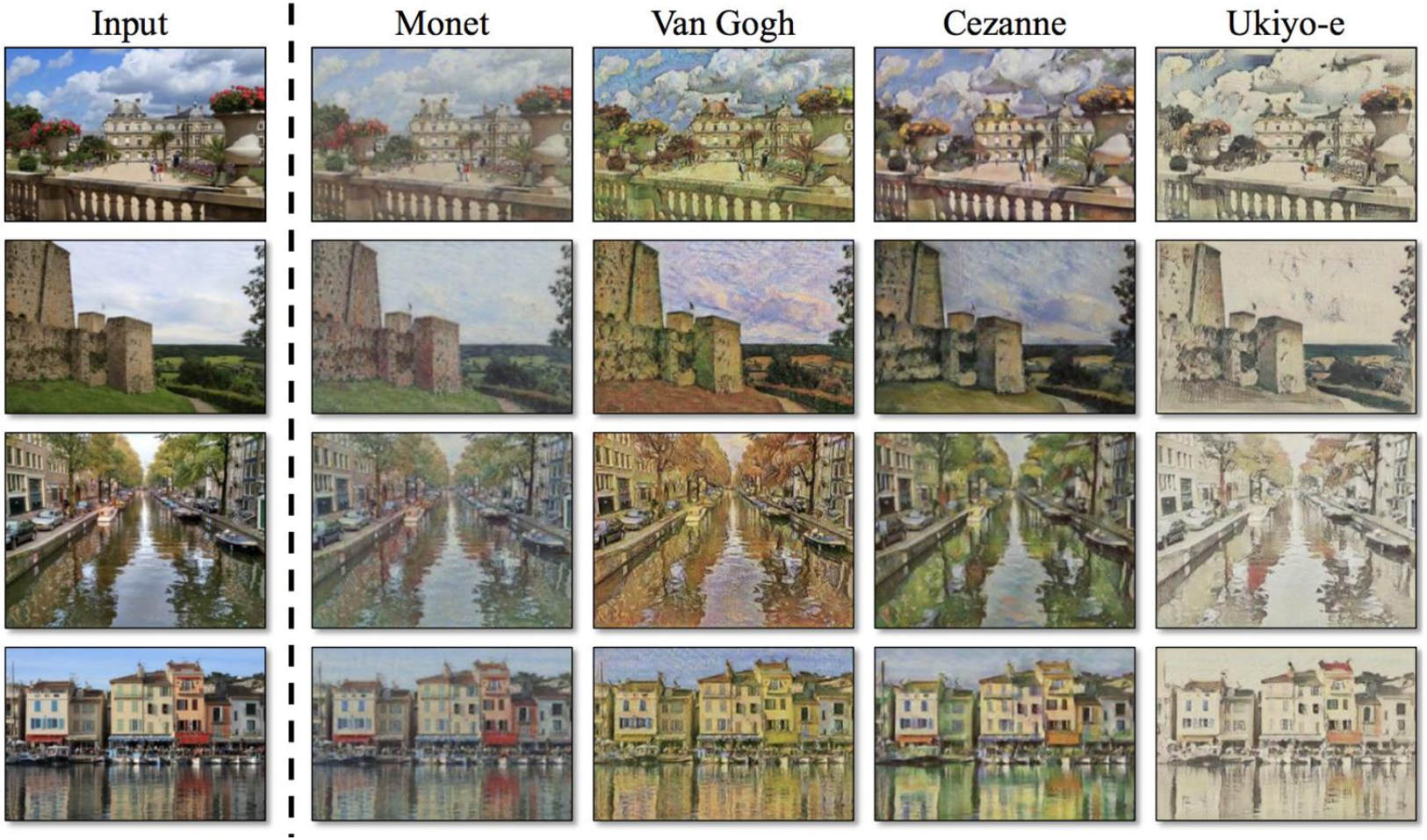
Loss	Per-pixel acc.	Per-class acc.	Class IOU
Cycle alone	0.22	0.07	0.02
GAN alone	0.51	0.11	0.08
GAN + forward cycle	**0.55**	**0.18**	**0.12**
GAN + backward cycle	0.39	0.14	0.06
CycleGAN (ours)	0.52	0.17	0.11

Table 4: Ablation study: FCN-scores for different variants of our method, evaluated on Cityscapes labels→photo.

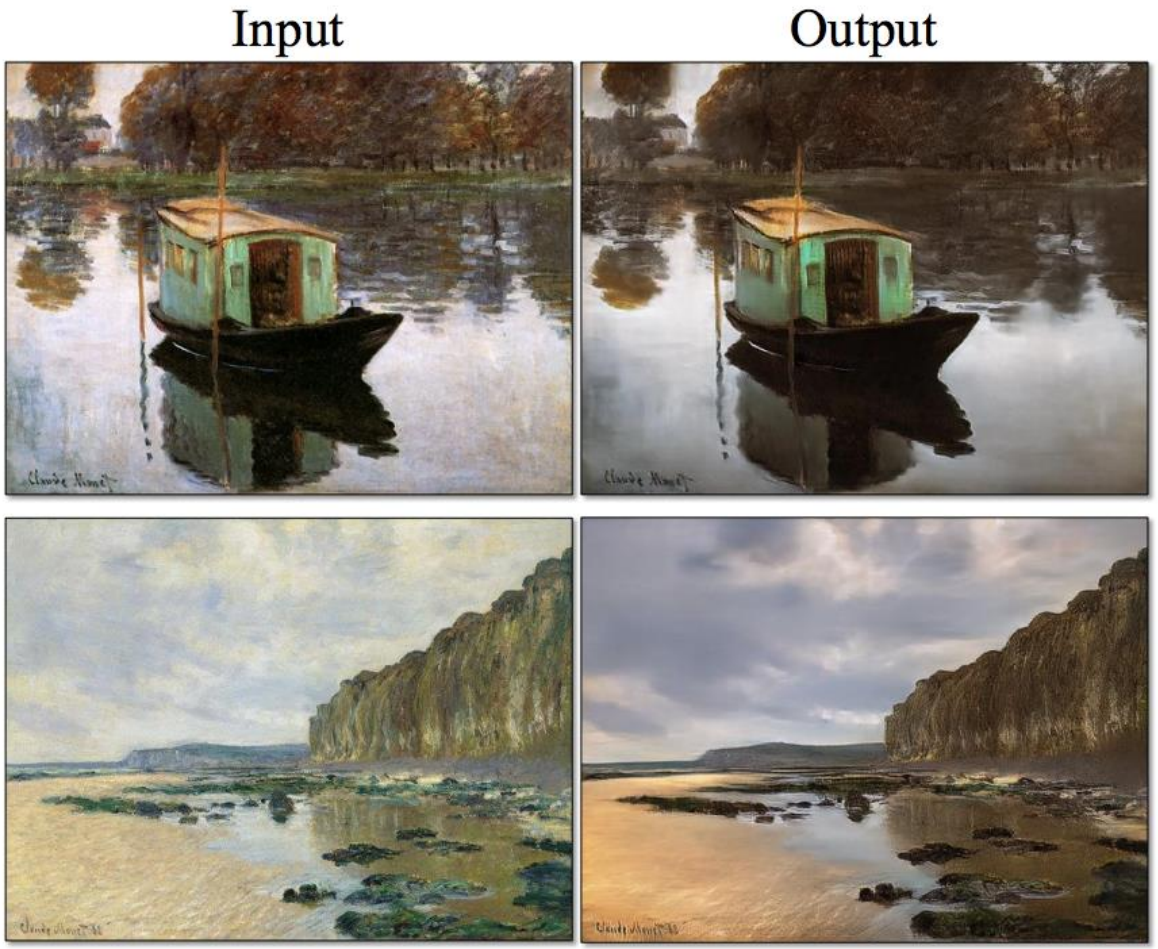
Loss	Per-pixel acc.	Per-class acc.	Class IOU
Cycle alone	0.10	0.05	0.02
GAN alone	0.53	0.11	0.07
GAN + forward cycle	0.49	0.11	0.07
GAN + backward cycle	0.01	0.06	0.01
CycleGAN (ours)	**0.58**	**0.22**	**0.16**

Table 5: Ablation study: classification performance of photo→labels for different losses, evaluated on Cityscapes.

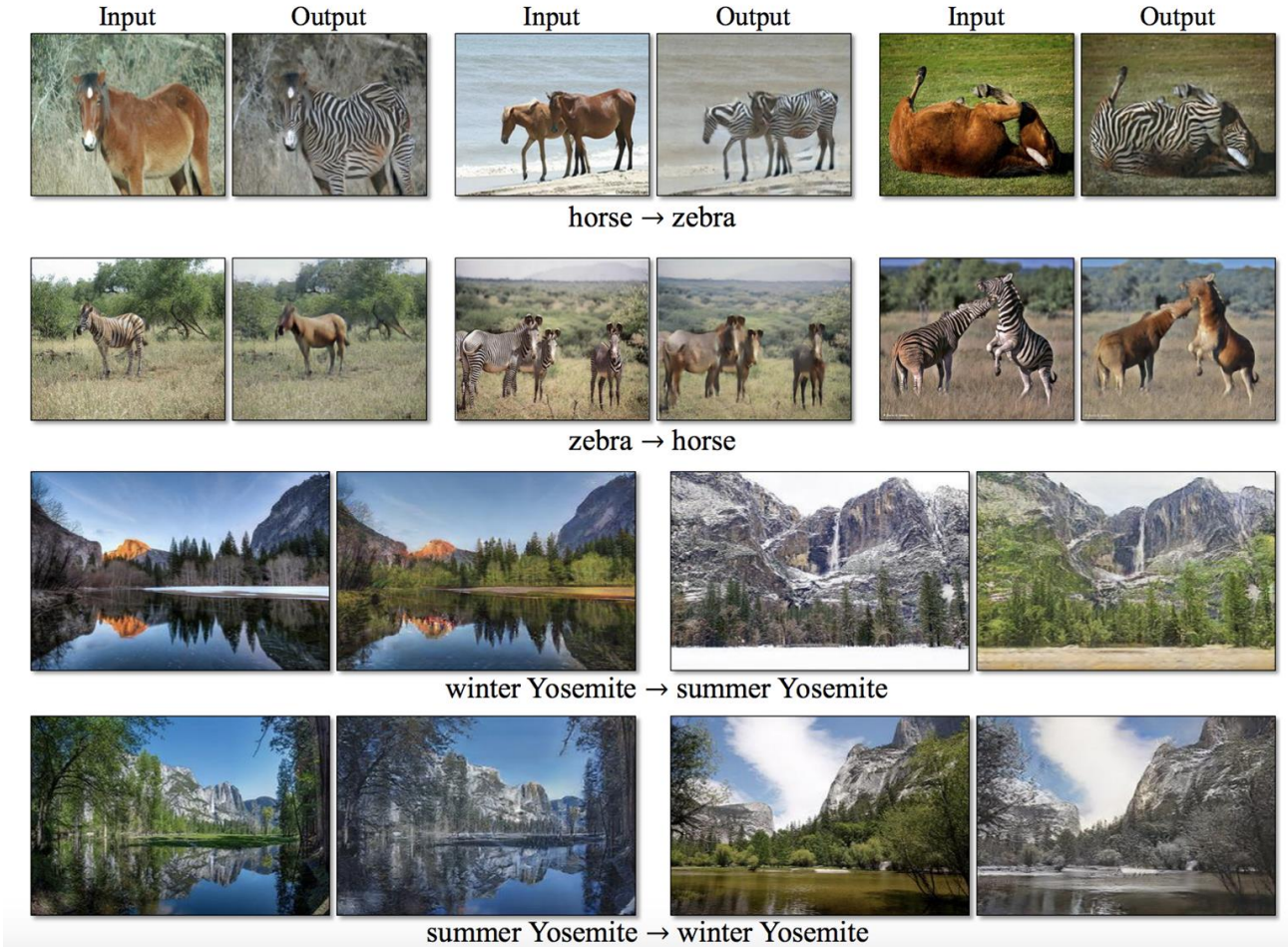
Experiments and Analysis



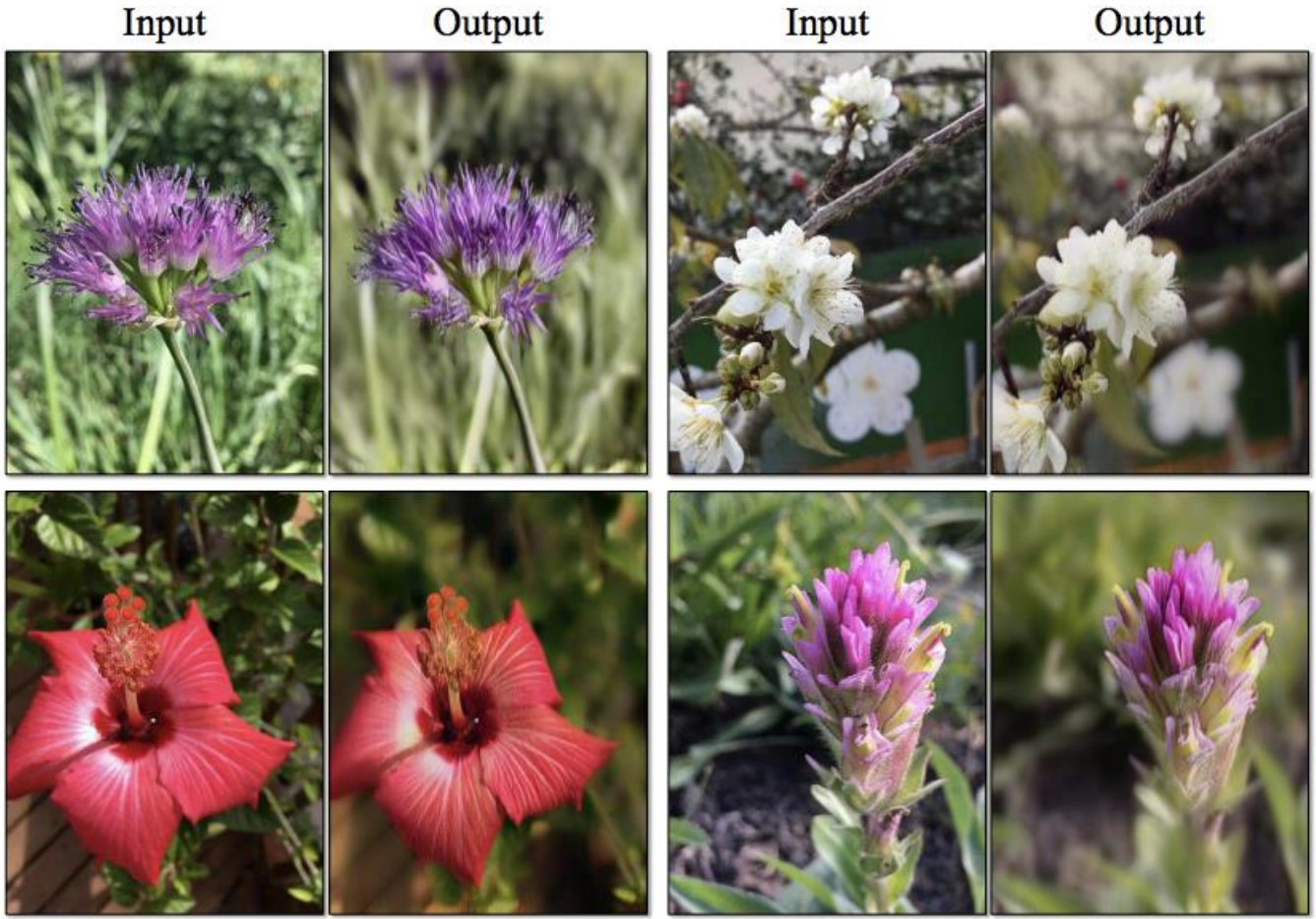
Unpaired image-to-Image Translation, Jun-Yan Zhu, Taesung Park, Phillip Isola, Alexei A. Efros (Feb 2018)



Unpaired image-to-Image Translation, Jun-Yan Zhu, Taesung Park, Phillip Isola, Alexei A. Efros (Feb 2018)



Unpaired image-to-Image Translation, Jun-Yan Zhu, Taesung Park, Phillip Isola, Alexei A. Efros (Feb 2018)

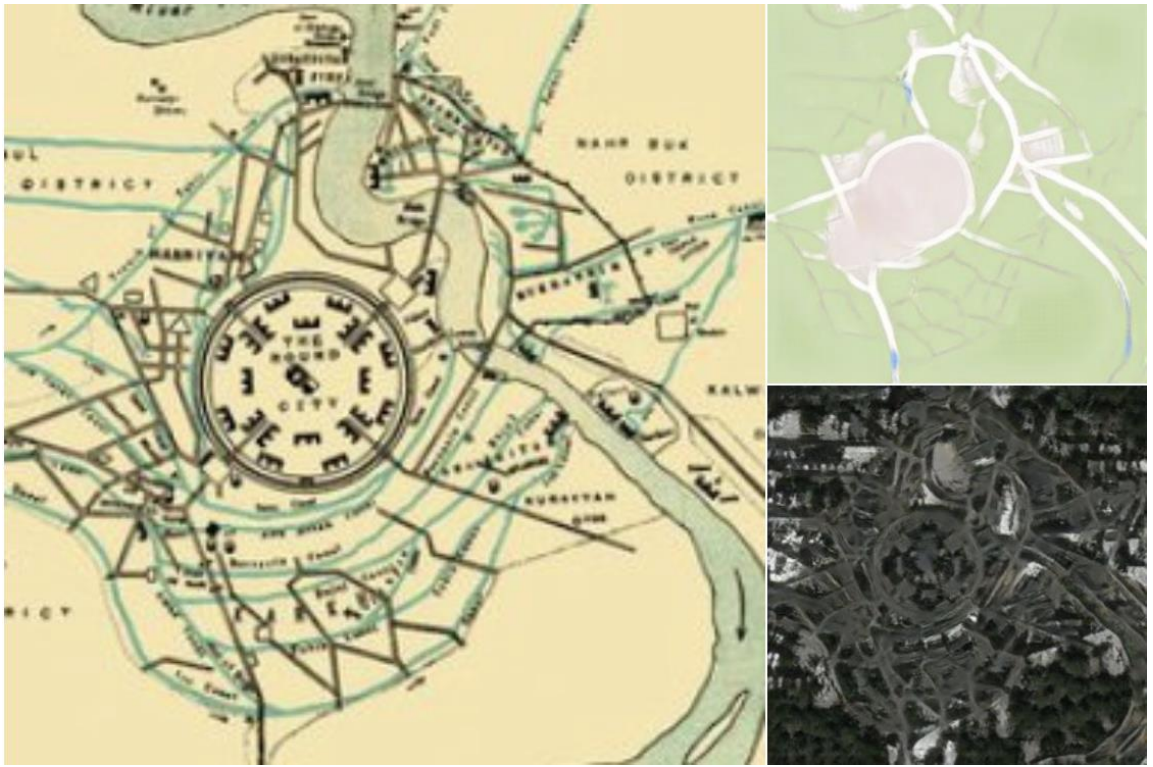


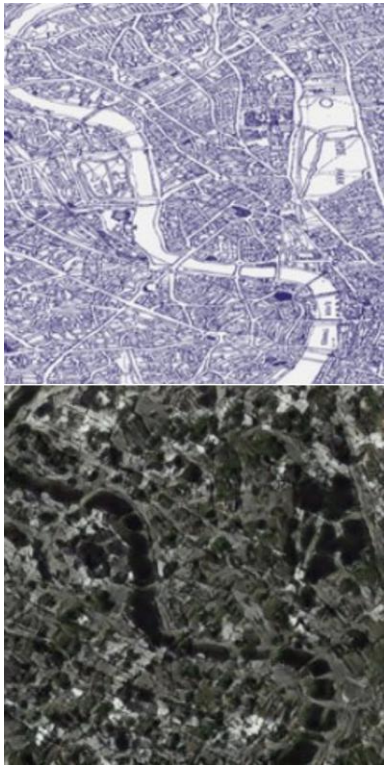
Unpaired image-to-Image Translation, Jun-Yan Zhu, Taesung Park, Phillip Isola, Alexei A. Efros (Feb 2018)

Face to Ramen

Unpaired image-to-Image Translation, Jun-Yan Zhu, Taesung Park, Phillip Isola, Alexei A. Efros (Feb 2018)

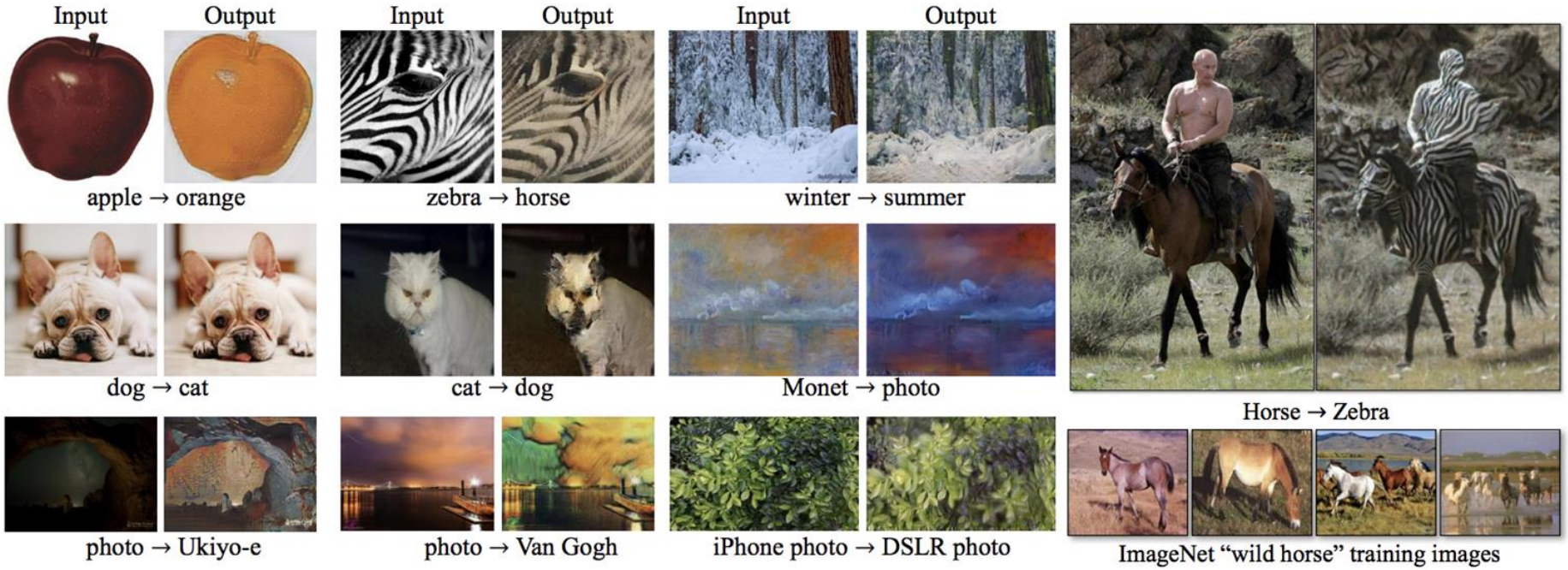
Resurrecting Ancient Cities





Unpaired image-to-Image Translation, Jun-Yan Zhu, Taesung Park, Phillip Isola, Alexei A. Efros (Feb 2018)

Limitations



Unpaired image-to-Image Translation, Jun-Yan Zhu, Taesung Park, Phillip Isola, Alexei A. Efros (Feb 2018)

References

- Conditional Gan, Mehdi Mirza and Simon Osindero (Nov 2014)
- Image-to-Image Translation, Philip Isola, Jun-Yan Zhu, Tinghui Zhou, Alexei A. Efros (Nov 2017)
- Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks Jun-Yan Zhu, Taesung Park, Phillip Isola, Alexei A. Efros (Feb 2018)