Memory Drawings-to-Scenes Translation

By Ran Xu and Ella Dagan

Image-to-Image Translation with Conditional Adversarial Networks

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

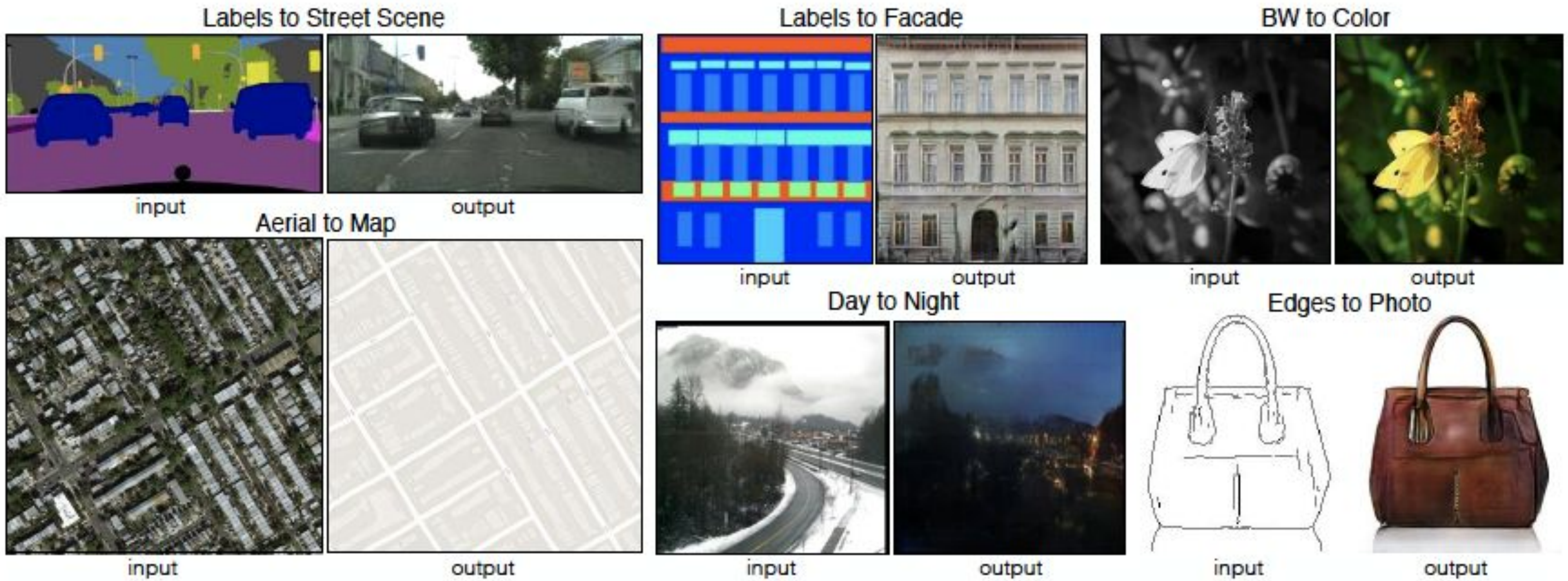


Figure 1: Many problems in image processing, graphics, and vision involve translating an input image into a corresponding output image. These problems are often treated with application-specific algorithms, even though the setting is always the same: map pixels to pixels. Conditional adversarial nets are a general-purpose solution that appears to work well on a wide variety of these problems. Here we show results of the method on several. In each case we use the same architecture and objective, and simply train on different data.

Conditional Adversarial Networks

- **Generative Adversarial Network (GANs):**
  - Generator and discriminator
  - Noise as input to generator
- **Conditional GANs (cGANs) :**
  - learn a conditional generative model as well as discriminator
  - noise and another condition (such as an observed image) as input to generator

Generator

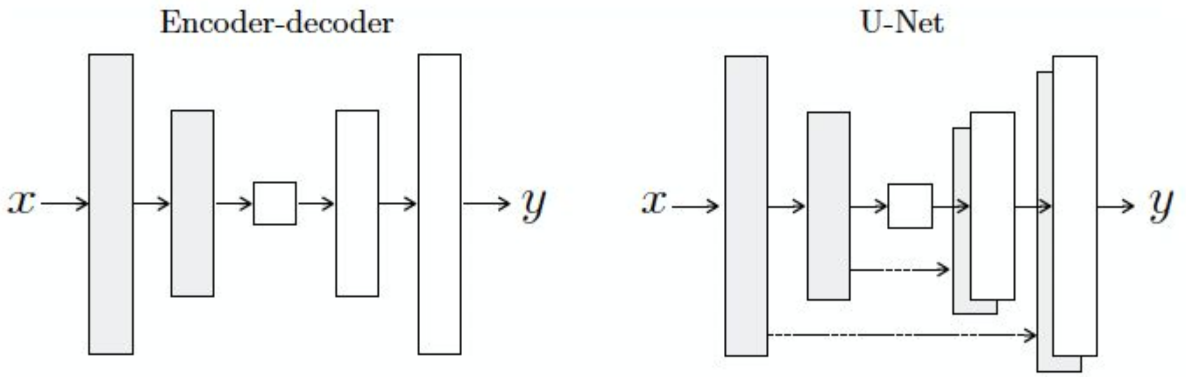


Figure 3: Two choices for the architecture of the generator. The "U-Net" [50] is an encoder-decoder with skip connections between mirrored layers in the encoder and decoder stacks.

U-Net: low level information reserved, higher resolution

Discriminator

- Compare generated images and ground truth, and decide whether the generated image is “real”
- Loss function: combining patchGAN and L1
  - L1 is better for low frequency correctness
  - patchGAN trying to classify whether a N*N patch from the image is real or fake (N is much smaller than the image size)
  - patchGAN is better for high frequency correctness

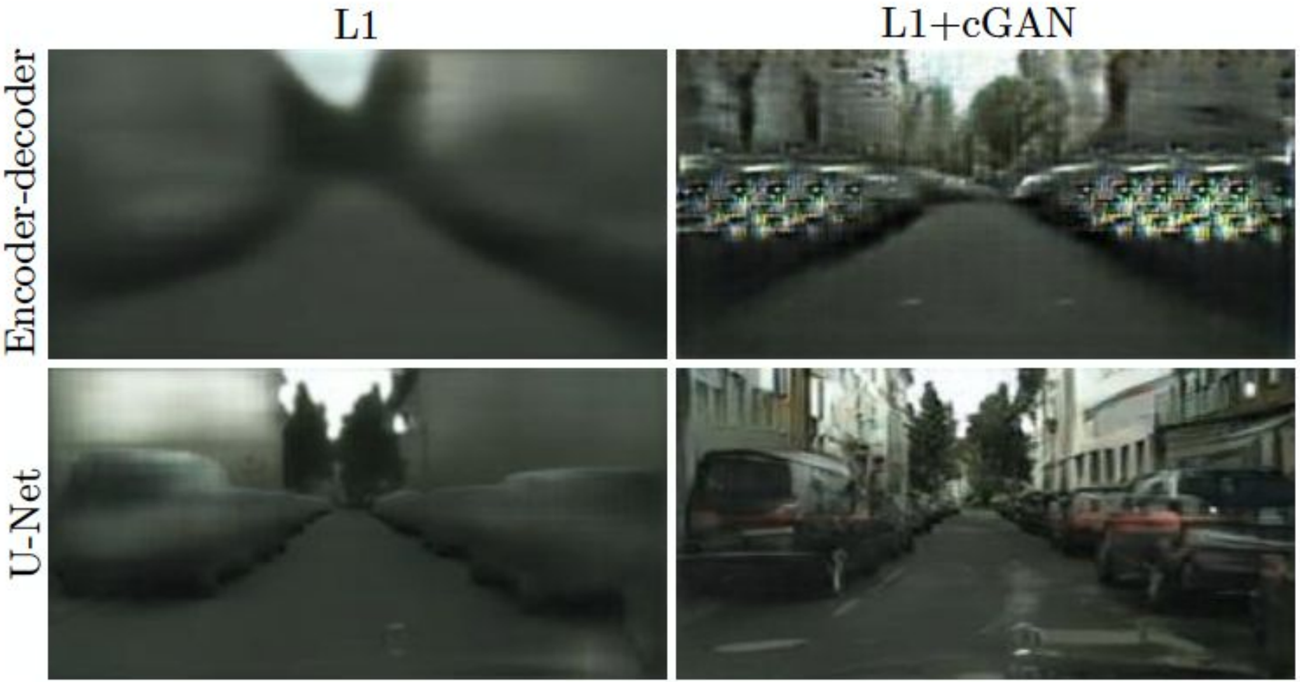


Figure 5: Adding skip connections to an encoder-decoder to create a "U-Net" results in much higher quality results.

Figure 4: Different losses induce different quality of results. Each column shows results trained under a different loss. Please see `https://phillipi.github.io/pix2pix/` for additional examples.

Loss	**Per-pixel acc.**	**Per-class acc.**	**Class IOU**
L1	0.42	0.15	0.11
GAN	0.22	0.05	0.01
cGAN	0.57	0.22	0.16
L1+GAN	0.64	0.20	0.15
L1+cGAN	**0.66**	**0.23**	**0.17**
Ground truth	0.80	0.26	0.21

Table 1: FCN-scores for different losses, evaluated on Cityscapes labels$\leftrightarrow$photos.

Figure 6: Patch size variations. Uncertainty in the output manifests itself differently for different loss functions. Uncertain regions become blurry and desaturated under L1. The 1x1 PixelGAN encourages greater color diversity but has no effect on spatial statistics. The 16x16 PatchGAN creates locally sharp results, but also leads to tiling artifacts beyond the scale it can observe. The 70×70 PatchGAN forces outputs that are sharp, even if incorrect, in both the spatial and spectral (colorfulness) dimensions. The full 286×286 ImageGAN produces results that are visually similar to the 70×70 PatchGAN, but somewhat lower quality according to our FCN-score metric (Table 3). Please see `https://phillipi.github.io/pix2pix/` for additional examples.

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

Table 3: FCN-scores for different receptive field sizes of the discriminator, evaluated on Cityscapes labels→photos. Note that input images are 256×256 pixels and larger receptive fields are padded with zeros.

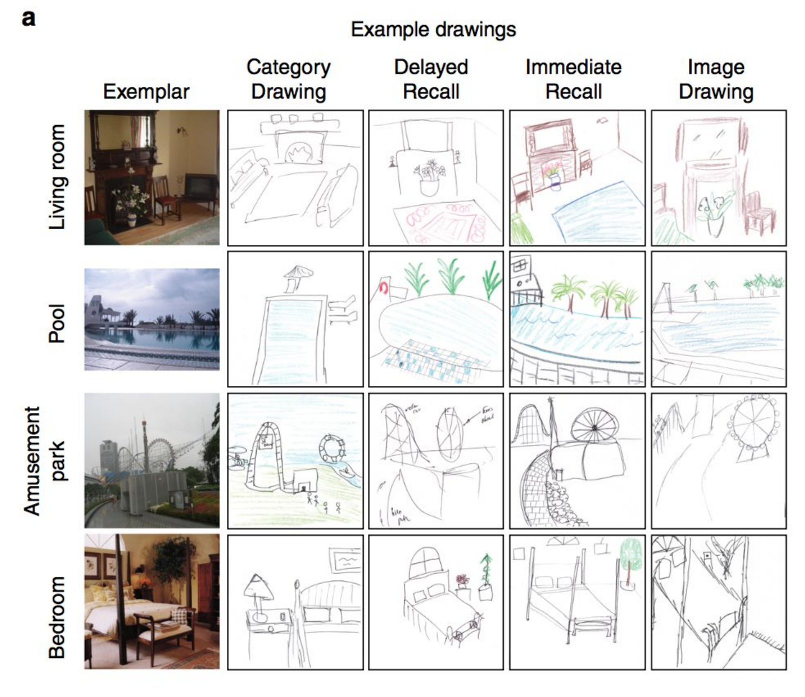
The Study

- Participants studied 30 scene categories (low & high memorability)
- 4 main conditions
- Produced 2682 drawings

Drawings of real-world scenes during free recall reveal detailed object and spatial information in memory

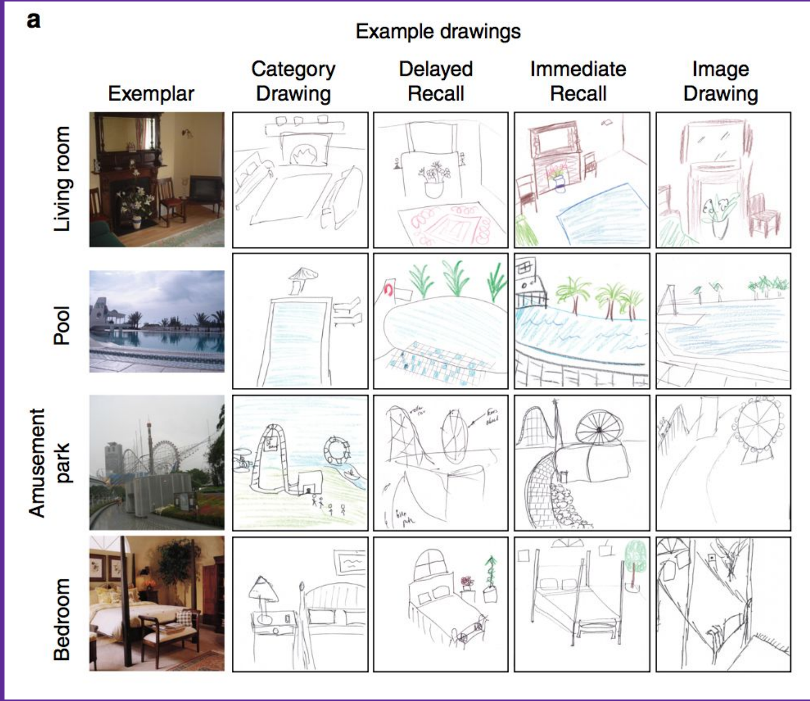
Wilma A. Bainbridge1, Elizabeth H. Hall1 & Chris I. Baker

https://doi.org/10.1038/s41467-018-07830-6



Main Results

- 2682 drawings quantified by AMT (number of objects, extraneous, spatial, size)
- Accurate spatial map of entire images
- Objects extremely close to original location



Discussion Questions -

Discussion Questions -

- What would you be interested in generating from the dataset mentioned in Bainbridge's article?

- They found that "drawing from memory reveal the object and spatial information maintained." Think about how human memory is different from a computer's memory, do you think that we can train a computer model to extract similar information from a scene? How will we do that?

Discussion Questions -

- What do you imagine the model would generate if we input a scene drawing then output a scene image, and then ask a person to draw that scene and input it as a new drawing. What scene would that image be like?