GAN Architectures for Domain Transfer in Image and Video

Peter Banyas, Zachary Charlick,
Ryan Ringel, Michael Scutari

Duke ENGINEERING

Overview

Generative Adversarial Networks (GANs):

- Generator creates new images.
- Discriminator discerns real from fake.

They learn in parallel, which puts pressure on the Generator to be more convincing.

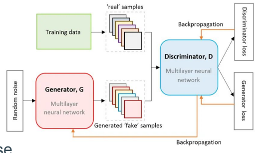

Figure 1: GANs use adversarial loss, which trains a generator and a discriminator. The generator proposes designs and the discriminator tries to tell them apart.
(Chandrahass TVS, "Exploring the Power of Generative Adversarial Networks (GANs) with Azure," Technical Blog, May 27, 2024, presidio).

Image-to-Image Translation:

For the input to the Generator: replace noise with real images from domain A.

Differences between Models:

- **Datasets** | Paired (**Pix**) vs Unpaired (**Cyc, Recyc**).
- **Loss** | Pixel-wise alignment: 1-way (**Pix**), 2-way (**Cyc**), & temporal (**Recyc**).
- **Architectures** | U-Net (**all**), Patch (**Pix, Cyc**), MultiScale & ResNet (**Recyc**).

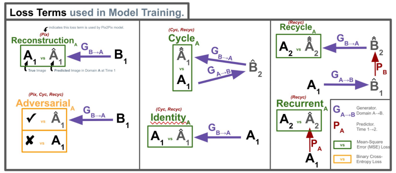

Figure 2: An index of the loss terms we used. Includes the standard adversarial loss that is common to all GANs, as well as the unique losses introduced by CycleGAN and R==ecycleGAN.

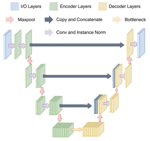

Figure 3: A U-Net breaks large images into smaller ones with deep channels, before re-expanding them back to their original size. This concentrates feature extraction, while preserving spacial information via skip connections.

Pix2Pix

U-Net Generator (+ **Attention**)
+ **PatchGAN Discriminator**.

Uses **paired** datasets ∴ it sees the "answer" to each "question" in training.

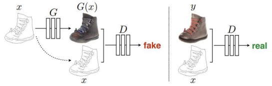

Figure 4: An overview of pix2pix Generator and Discriminator models (From P. Isola, 2018)

Reconstruction loss pushes Gen to learn pixel-wise transform across domains.

Adversarial loss pushes Discr to detect oddities in fake images.

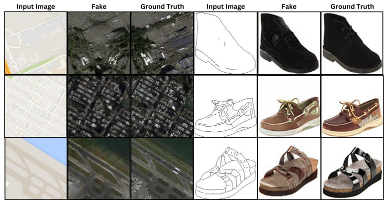

Figure 5: Pix2pix results pictured for map2sat (left) and sketch2shoe (right) paired datasets. "Fake" images are generated from the trained pix2pix models

CycleGAN

U-Net Generator + **PatchGAN** Discriminator.

Visible Artifacts suggest blind pattern application.

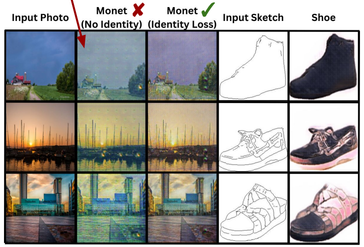

Figure 6: CycleGAN results pictured for photo2monet (left) and sketch2shoe (right) unpaired datasets. Photo2monet was tested with and without identity loss

Uses **unpaired** datasets
∴ learns thematic differences, not direct transforms.

Cycle loss pushes Gen to preserve features through translation.
➔ It must be able translate backwards & restore original.

Identity loss pushes Gen to preserve features that already match output theme.
➔ It must not change inputs that also fit output domain.

RecycleGAN

Figure 7. results of cyclegan trained on lion king movie frames

GOAL: Extend Image Translation to Videos.

U-Net Generator *(+shallower)*
+ **Multi-Scale Patch** Discriminator *(+downsampling)*
+ **ResNet Predictor** *(translates image into future)*

Recycle & **Recurrent** loss push temporal consistency.
➔ Only Pred can change time; Gen must preserve it.

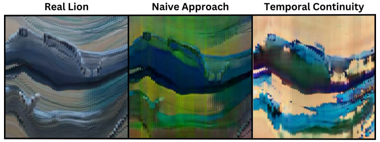

Figure 8. Temporal continuity of horizontal video slices shown for both approaches

Figure 9. Eight sequential frames of the ground truth model, naive approach using cycleGAN, and temporal continuity approach using recycleGAN

Evaluation Metrics

Structural Similarity Index (SSIM)

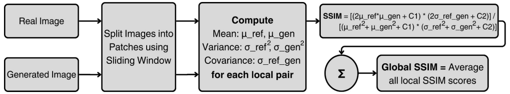

Figure 10: We use SSIM to compare images when ground truth is available. This is possible for models trained with paired (Pix2Pix) or quasi-paired data sets (CycleGAN).

SSIM compares **luminance**, **contrast**, and **structure** on windows of two images, making it very useful for evaluating **paired** data sets where ground truth is available.

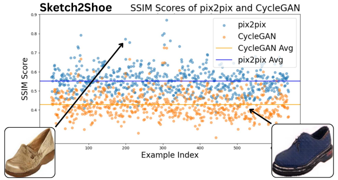

Figure 11: Results from our SSIM scores of models trained using the sketch2shoe dataset align with our qualitative intuition about the generated image quality. Our pix2pix-generated model consistently outperforms CycleGAN for this task. However, there is insufficient data to conclude whether this is due to the architecture itself or to hyperparameters and other network tuning.

Fréchet Inception Distance (FID)

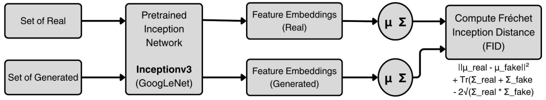

Figure 12: FID computes the feature embeddings using a pretrained inception network. For our project we used the Inceptionv3 model from GoogLeNet. The mean and covariance of each set of features are computed and compared using a modified Wasserstein Distance.

FID evaluates **unpaired** datasets by comparing distributions of generated and real images using a trained **inception network**.

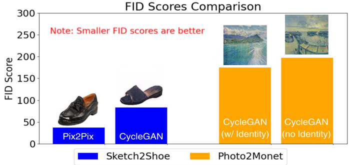

Figure 13: FID is necessary when a ground truth image is not available in the case of unpaired data sets (CycleGAN). We can also evaluate paired models by unpairing the images into sets during FID computation. We found that our Pix2Pix model outperforms our CycleGAN for Sketch2Shoe, and that CycleGAN with identity provides better results for the Photo2Monet model

Abridged References

P. Isola, J.-Y. Zhu, T. Zhou, A. Efros (2018). "Image-to-Image Translation with Conditional Adversarial Networks." Available: Available: https://arxiv.org/pdf/1611.07004

Jun-Yan Zhu, Taesung Park, Phillip Isola, Alexei A. Efros (2020). "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks" Available: https://arxiv.org/abs/1703.10593

A. Bansal, S. Ma, D. Ramanan, Y. Sheikh (2018). "Recycle-GAN: Unsupervised Video Retargeting". In: ECCV.

Zhou Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," In: *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600-612, April 2004