# Face-to-Sketch Translation Using Cycle-Consistent Adversarial Networks

**Pavel Kratochvíl, Jakub Kuzník and Lucie Svobodová**

xkrato61@stud.fit.vutbr.cz, xkuzni04@stud.fit.vutbr.cz, xsvobo1x@stud.fit.vutbr.cz

## Abstract

Utilizing various convolutional neural nets (CNNs) for image-to-image translation has shown promising results. However, these netowrks typically rely on paired sets of labeled images, which can be challenging to obtain. Cycle-Consistent Generative Adversarial Network (CycleGAN) offers a solution by taking advantage of unpaired sets of images. Nonetheless, CycleGANs face challenges in effectively accommodating variations in geometric shapes and textures, as well as accurately capturing important features within the images. In this work, we aim to enhance CycleGAN for the task of Face-to-Sketch translation. To achieve this, we will try to utilize AdaLIN normalization to address variations in shapes and textures, and employ the Self-Attention mechanism to extract important features from input images.

## 1 Introduction

Years of research in computer vision and image processing produced powerful translation systems (e.g.,[1], [2], [3], [4], [5], [6], [7], [8], [9], [10]); however, these were usually supervised, relying on available example image pairs, which are often difficult to obtain.

The Cycle-Consistent Adversarial Network (CycleGAN) was introduced in [11] for image-to-image translation. One of its notable advantages, in contrast to alternative methodologies, is its unsupervised nature, eliminating the necessity for paired input-output training examples .

CycleGAN captures the distinctive features of one image collection and determines how these features can be translated into another image collection, facilitating tasks such as transferring images of horses to images of zebras and vice versa [11].

This problem broadly concerns image-to-image translation, encompassing translations such as image to sketches [12]. Essentially, it involves converting an image from one representation of a given scene, denoted as x, to another representation, denoted as y, such as grayscale to color, or edge-map to photograph.

While supervision in the form of paired examples may be lacking, we can leverage supervision at the set level: given one set of images in domain $X$ and another set in domain $Y$, we can train a mapping function $G : X \rightarrow Y$. This function produces outputs $\hat{y} = G(x)$, where $x \in X$, indistinguishable from images $y \in Y$ by an adversary trained to differentiate $\hat{y}$ from $y$. Theoretically, this objective can induce an output distribution over $\hat{y}$ that aligns with the empirical distribution $p_{\text{data}}(y)$ (typically requiring $G$ to be stochastic). Consequently, the optimal $G$ effectively translates the domain $X$ to a domain $Y'$ distributed identically to $Y$ [12].

Considerable research has been conducted in the area of sketch-to-image conversion (e.g., [13], [14] [15]), yielding promising results. However, these studies often present a limited subset of their outcomes, typically showcasing only the most favorable examples.

In this study, we employ CycleGAN across various datasets with a specific aim: to translate images of real people's faces into sketches and vice versa. Leveraging CycleGAN's unsupervised nature, we endeavor to effectively capture distinct facial features and faithfully transpose them between photographic and sketched representations.

## 2 The Original Work

Zhu et al. established seminal groundwork in the field of CycleGAN through their publication on unpaired image-to-image translation [11]. This work addresses the scenario wherein training data lacks explicit correspondences, comprising a source set $\{x_i\}_{i=1}^{N}$ ($x_i \in X$) and a target set $\{y_j\}_{j=1}^{M}$ ($y_j \in Y$), where no explicit pairing information between elements of $X$ and $Y$ is provided.

However, the work assumes that there is some underlying relationship between the domains, though not explicitly delineated, providing a foundational premise for the subsequent development of CycleGAN methodologies.

The aim is to train a mapping function $G : X \rightarrow Y$, such that the output $\hat{y} = G(x)$, $x \in X$, is unrecognizable from images $y \in Y$ by an adversary trained to distinguish $\hat{y}$ apart from $y$. However, such a translation does not guarantee a meaningful correspondence between $x$ and $y$. This limitation led to the introduction of "cycle consistency," whereby if, for instance, a face image is translated to a sketch and then translated back, the result should revert to

the original image. Mathematically, if we have a translator $G : X \to Y$ and another translator $F : Y \to X$, then $G$ and $F$ should be inverses of each other, as depicted in Figure 1.

This is achieved by utilizing two discriminators and two generators, with each generator being paired with a discriminator. Each discriminator is trained to assess whether the translated images are indistinguishable from the other images in the actual domain. The generated picture should be indistinguishable from the original one after completing the entire cycle.

In this study, two distinct techniques are employed: adversarial losses [16] and cycle consistency loss. Adversarial losses aimed to aligning the distribution of generated images with the data distribution in the target domain, and cycle consistency losses, which serve to ensure the coherence of the learned mappings $G$ and $F$. Adversarial losses [16] are utilized for both mapping functions. Specifically, for the mapping function $G : X \to Y$ and its discriminator $D_Y$, the objective is formulated as follows:

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

Where $G$ tries to generate images $G(x)$ that look similar to images from domain $Y$, while $D_Y$ aims to distinguish between translated samples $G(x)$ and real samples $y$. $G$ aims to minimize this objective against an adversary $D$ that tries to maximize it, i.e., $\min_G \max_{D_Y} \mathcal{L}_{\text{GAN}}(G, D_Y, X, Y)$ [11]. The similar adversarial loss is introduced for the mapping function $F : Y \to X$ and its discriminator $D_X$.

For each image $x$ from domain $X$, the image translation cycle should be able to bring $x$ back to the original image, i.e., $x \to G(x) \to F(G(x)) \approx x$. We call this forward cycle consistency. Similarly, for each image $y$ from domain $Y$, $G$ and $F$ should also satisfy backward cycle consistency: $y \to F(y) \to G(F(y)) \approx y$. This behavior is achieved using a cycle consistency loss:

$$\begin{aligned}\mathcal{L}_{\text{cyc}}(G, F) = & \mathbb{E}_{x \sim p_{\text{data}}(x)}\left[\|F(G(x)) - x\|_1\right] \\ &+ \mathbb{E}_{y \sim p_{\text{data}}(y)}\left[\|G(F(y)) - y\|_1\right].\end{aligned} \tag{1}$$

That means that the full objective of the problem is given by:

$$\begin{aligned}\mathcal{L}(G, F, D_X, D_Y) = & \mathcal{L}_{\text{GAN}}(G, D_Y, X, Y) \\ &+ \mathcal{L}_{\text{GAN}}(F, D_X, Y, X) \\ &+ \lambda \mathcal{L}_{\text{cyc}}(G, F),\end{aligned} \tag{2}$$

where $\lambda$ controls the relative importance of the two objectives and the aim is to solve:

$$G^*, F^* = \arg\min_{G,F} \max_{D_X, D_Y} \mathcal{L}(G, F, D_X, D_Y). \tag{3}$$

## 3 Baseline Solution

The baseline solution presented in this work is implemented based on the original CycleGAN paper by Zhu et al. [11]. As a foundation for our implementation, we used an implementation by Aladdin Persson[1]. We have made some changes to closely resemble the architecture and methodology described in the original paper. The implementation is developed in Python using the PyTorch framework[2].

The baseline solution comprises four main components: two generators and two discriminators.

Each generator's architecture utilizes a $128x128$ input dimension, where a series of three residual convolutional layers with pooling is employed to effectively extract features. Following each convolutional operation, rectified linear unit (ReLU) activation functions are applied. The use of residual connections ensures that information and features are preserved and propagated deeper into the network, mitigating the risk of information loss.

Subsequently, two fractionally-strided convolutional layers, each with a stride of $\frac{1}{2}$, are employed. These layers serve to upscale the feature maps, enhancing spatial resolution while maintaining the integrity of extracted features.

Finally, a single convolutional layer is employed to extract RGB features, culminating in the generation of the output image. This cascade of operations adheres to established principles in convolutional neural network design, ensuring efficient feature extraction and preservation throughout the network architecture

The discriminator networks are entrusted with the responsibility of discriminating between translated images and authentic images from the target domain. Analogous to the generators, discriminators utilize three convolutional layers with Instance Normalization applied after each layer. However, in this configuration, the convolutional layers are not residual, and LeakyReLU activation functions are employed instead of ReLU to ensure that even negative information is propagated through the network. Finally, a convolutional layer with a sigmoid activation function is appended to determine the credibility of the generated image.

The training procedure follows the adversarial training framework outlined in the original CycleGAN paper. We utilize adversarial losses to align the distribution of generated images with the data distribution in the target domain. Additionally, we incorporate cycle consistency loss to enforce the mapping functions' coherence and ensure that translated images maintain meaningful correspondence with their original counterparts.

During training, we employ optimization techniques

[1] Aladdin Persson's repository: https://github.com/aladdinpersson/Machine-Learning-Collection/

[2] PyTorch: https://pytorch.org

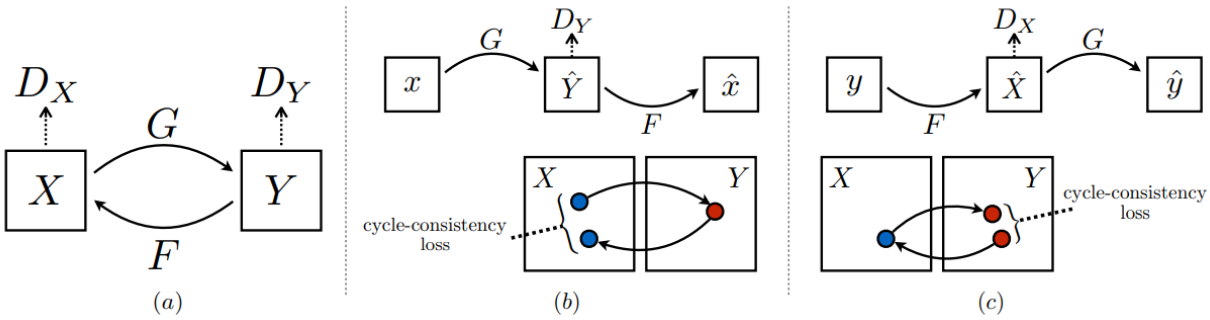


Figure 1. The CycleGAN model consists of two mapping functions: $G : X \rightarrow Y$ and $F : Y \rightarrow X$, alongside associated adversarial discriminators $D_Y$ and $D_X$. These discriminators encourage $G$ to translate $X$ into outputs indistinguishable from domain $Y$, and similarly for $D_X$ and $F$. To further regularize the mappings, two cycle consistency losses are introduced. These losses ensure that translating from one domain to the other and back again leads to the original image. The forward cycle-consistency loss (b) is defined as $x \rightarrow G(x) \rightarrow F(G(x)) \approx x$, and the backward cycle-consistency loss (c) is defined as $y \rightarrow F(y) \rightarrow G(F(y)) \approx y$ [11].

such as stochastic gradient descent (SGD) or Adam to update the parameters of the generator and discriminator networks iteratively. Hyperparameters such as learning rate, batch size, and weight decay are carefully tuned to facilitate efficient convergence and stable training.

## 4 Baseline Solution Evaluation

While the baseline CycleGAN solution demonstrates promising capabilities in certain scenarios, it encounters significant challenges and limitations in others. Here, we discuss some of the notable problems encountered during experimentation and highlight key observations from the generated images. The experiments were conducted using datasets described in Section 6.

**Generating Sketches from Photos**

One area where the baseline solution performs relatively well is in generating sketches from real photos. Example (a) in Figure 2 showcases a generated sketch from an original photo at around 15,400 iterations. The generated sketch resembles the original photo, capturing essential features and details, but in a simplified black-and-white format. Similarly, when generating anime-style sketches from real photos, it also performs very well as can be seen in image pair (e) in Figure 2 after only 40,000 iterations of training.

**Generating Photos from Sketches**

However, when it comes to generating photos from sketches, the baseline solution encounters significant difficulties. Example (b) in Figure 2 illustrates a generated photo from an original sketch at approximately 181,000 iterations, exhibiting poor quality and lacking resemblance to any real face. With prolonged training, the quality of generated photos from sketches further deteriorates, as depicted in example (c) after 278,600 iterations. Instead of representing a real face, it exhibits arbitrary color changes without retaining the underlying structure of the sketch. This degradation suggests a limitation when the model focuses too much on the backgrounds and not the main part of the image, the face itself. A similar challenge is observed when generating photos from anime-style sketches, as shown in example (d) after 140,000 iterations. Despite prolonged training, the model struggles to produce satisfactory results and tends to exhibit similar issues as described above.

**Generating Simple Sketches from Photos**

Generating realistic images from simple sketches of, for example, mushrooms, poses challenges due to the limited information in the sketches. Example (i) in Figure 2 demonstrates this issue, where the model tends to produce empty images when translating from photos of mushrooms to sketches.

**Translation Between Photos of Different Species**

When translating between horse and zebra images, where minimal structural changes are required, the baseline solution succeeds due to the similarity in overall structure and texture, as seen in example (f) in Figure 2. Conversely, translating between images of diverse species like cats and dogs poses challenges due to significant differences in appearance and structural features, as depicted in example (g) in Figure 2.

## 5 Proposed Improvements

In this section, we propose several targeted enhancements to the baseline CycleGAN solution, aiming to overcome specific limitations and improve performance in key areas. Notably, we aim to address the disparity in perfor-

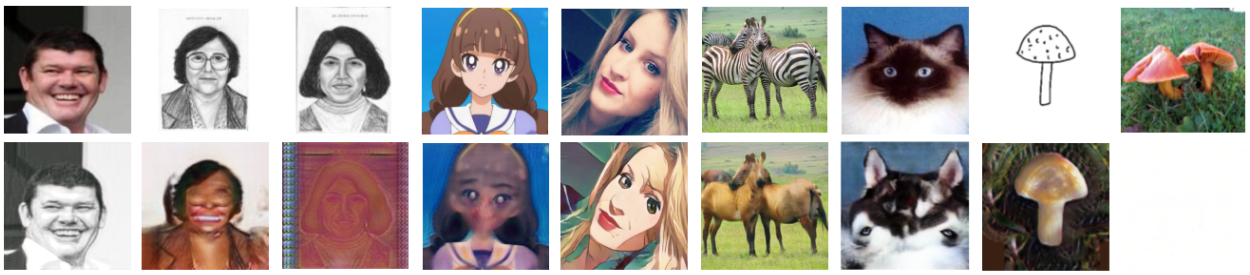


Figure 2. Evaluation of the baseline CycleGAN solution on various datasets described in Section 6. Original images are displayed in the top row, images generated from them are displayed in the bottom row. Examples (a), (b), and (c) utilize the CelebA and CUFS datasets. (a) represents a sketch generated after 15,400 iterations, (b) and (c) displays real faces generated after 181,000 and 278,600 iterations, respectively. For (d) and (e), the selfie2anime dataset is used, with (d) showing a real face generated from an anime sketch after 140,000 iterations, and (e) displaying an anime sketch generated from a real face after 40,000 iterations. Example (f) illustrates the translation from zebra to horse, and (g) demonstrates cat to dog translation, both taken from [13]. Examples (h) and (i) shows experiments conducted on the Sketchy database with (h) showing a photo generated from a sketch, and (i) showcasing a sketch generated from a photo, both after 40,000 iterations.

mance observed in generating real photos from sketches compared to generating sketches from real photos. We would like to improve the model's ability to handle significant geometric changes, by exploring techniques that better preserve essential features and characteristics during translation.

**AdaLIN Normalization**

One of the challenges encountered in our solution is its inability to effectively adapt to changes in geometric structures as can be seen in Figure 2 (i). A potential solution for this issue was proposed in [13], where Adaptive Layer-Instance Normalization (AdaLIN) was introduced. This normalization technique serves as an alternative to Instance Normalization, which may struggle to capture diverse geometric shapes and textures adequately. The parameters of AdaLIN are learned during training on the dataset, allowing it to dynamically adjust the balance between Instance Normalization (IN), already in use, and Layer Normalization (LN). This adaptive approach offers improved flexibility in accommodating variations in geometric structures and textures within the data.

**Self-Attention**

Another issue arises, especially in the later phases of training, when the network tends to prioritize the background of the image over other features as can be seen in Figure 2 (i). This challenge is addressed by the Self-attention mechanism for Convolutional Networks proposed by Bello et al. [17]. While self-attention mechanisms are commonly used in sequence modeling, Bello et al. (2020) demonstrate their application in convolutional networks, where they are utilized to extract crucial parts of the image. The attention mechanism operates concurrently with the convolutional layers, and their outputs are combined in the final convolutional layer responsible for generating the resulting image. This integration ensures that the important features are highlighted with greater precision.

## 6 Datasets

For evaluating the performance of the models, we will assess them across a range of sketch styles, including anime-style and "realistic" sketches. Additionally, we will explore other image-to-image translation tasks, such as transforming images of horses into zebras, cats to dogs, or converting real photos of mushrooms to narrow one-line sketches and vice versa.

For experiments involving anime-style sketches, we will primarily use the selfie2anime[3] dataset, which consists of 3400 female selfie images and 3400 female character images for training, with an additional 100 images for validation. This dataset was used in [13].

To assess models' performance with more "realistic" sketches, we will combine the CelebA[4] and CUFS[5] datasets. Regarding the CUFS dataset, it comprises 188 paired images of university students and their corresponding sketches, along with additional 1194 unpaired sketches. To ensure a balanced dataset for training and validation, the unpaired sketches will be combined with randomly selected images from the CelebA dataset. This

[3]selfie2anime dataset: https://drive.google.com/file/d/1xOWj1UVgp6NKMT3HbPhBbtq2A4EDkghF/view

[4]CelebA dataset: https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html

[5]CUFS dataset: https://www.kaggle.com/datasets/arbazkhan971/cuhk-face-sketch-database-cufs

combined dataset will then be divided into a training set consisting of 1300 images and a validation set containing 83 images for each domain. The CUFS dataset was used in [18].

In addition to these datasets, we will also utilize the horse2zebra[6] dataset, consisting of 1187 horse and 1474 zebra images, as well as the cat2dog[7] dataset, which includes 871 images of cat faces and 1364 images of dog faces. These datasets have been previously used in studies on unpaired image-to-image translation[11, 13].

Furthermore, we will incorporate the Sketchy database[8], which contains sketch-photo pairs comprising 75,471 sketches of 12,500 real images spanning various domains, especially animals, plants, and everyday objects. This dataset has been utilized in research focusing on sketch-based image retrieval [19].

## 7 Future Training and Evaluation

Like other image generation tasks, unpaired image-to-image translation lacks good quantitative metrics, as traditional metrics may not be well-suited due to the lack of paired data for comparison. Metrics like PSNR, commonly used for image quality assessment, require paired data for accurate evaluation, rendering them unsuitable for tasks such as faces-to-sketches translation. We will try to utilize Inception Score[20] as a quantitative metric, but its applicability may be limited.

For qualitative evaluation, we conduct visual comparisons across a variety of images, spanning diverse subjects and scenarios. This comprehensive approach enables a thorough assessment of our models' performance across different image-to-image translation tasks.

The training of our models is conducted on Metacentrum[9] clusters, primarily utilizing the Zia and Adan clusters.

## 8 Conclusion

This work proposes enhancements to CycleGAN for Face-to-Sketch translation by incorporating AdaLIN normalization and Self-Attention mechanisms.

While CycleGAN has shown promise in various image-to-image translation tasks, it faces challenges in accurately capturing geometric shapes, textures, and important features. By integrating AdaLIN normalization, we aim to address variations in shapes and textures, while the Self-Attention mechanism seeks to extract crucial features from input images more effectively.

Through experimentation and evaluation, we anticipate that these enhancements will improve the model's performance in translating images of real faces into sketches and vice versa. Our proposed improvements aim to overcome the limitations observed in the baseline solution, particularly in generating realistic images from sketches. By refining the model's ability to handle geometric changes and prioritize essential features, we strive to enhance the accuracy of the translated images.

Moving forward, we plan to conduct thorough training and evaluation using diverse datasets, including anime-style sketches and more realistic sketches. We aim to contribute to the advancement of image-to-image translation techniques, particularly in the domain of Face-to-Sketch translation, through our proposed improvements to CycleGAN.

[6]horse2zebra dataset: `https://www.kaggle.com/datasets/balraj98/horse2zebra-dataset`

[7]cat2dog dataset: `https://www.kaggle.com/datasets/waifuai/cat2dog`

[8]Sketchy database: `https://faculty.cc.gatech.edu/~hays/tmp/sketchy-database.pdf`

[9]Metacentrum: `https://metavo.metacentrum.cz/cs/`

## References


[1] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *ICCV*, 2015.

[2] Aaron Hertzmann, Charles E. Jacobs, Nevin Oliver, Brian Curless, and David H. Salesin. Image analogies. In *SIGGRAPH*, 2001.

[3] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

[4] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016.

[5] Pierre-Yves Laffont, Zhile Ren, Xin Tao, Chen Qian, and James Hays. Transient attributes for high-level understanding and editing of outdoor scenes. *ACM TOG*, 33(4):149, 2014.

[6] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[7] Yale Shih, Sylvain Paris, Frédo Durand, and William T Freeman. Data-driven hallucination of different times of day from a single outdoor photo. *ACM TOG*, 32(6):200, 2013.

[8] Xiaolong Wang and Abhinav Gupta. Generative image modeling using style and structure adversarial networks. In *ECCV*, 2016.

[9] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *ICCV*, 2015.

[10] Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. In *ECCV*, 2016.

[11] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks, 2020.

[12] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017.

[13] Junho Kim, Minjae Kim, Hyeonwoo Kang, and Kwanghee Lee. U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation, 2020.

[14] Nastaran Moradzadeh Farid, Maryam Saeedi Fard, and Ahmad Nickabadi. Face sketch to photo translation using generative adversarial networks, 2021.

[15] Ramchandra Giri, Badri Lamichhane, and Biplove Pokhrel. Sketch to image translation using generative adversarial network. *Journal of Engineering and Sciences*, 2:70–75, 12 2023. doi:10.3126/jes2.v2i1.60397.

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014.

[17] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V. Le. Attention augmented convolutional networks, 2020.

[18] Wei Zhang, Xiaogang Wang, and Xiaoou Tang. Coupled information-theoretic encoding for face photo-sketch recognition. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 513–520. IEEE, 2011.

[19] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: learning to retrieve badly drawn bunnies. 35(4), jul 2016. ISSN 0730-0301. doi:10.1145/2897824.2925954. URL `https://doi.org/10.1145/2897824.2925954`.

[20] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans, 2016.