Benign Examples: Imperceptible Changes Can Enhance Image Translation Performance

Vignesh Srinivasan,[1] **Klaus-Robert Müller,**[2,3,4,5,*] **Wojciech Samek,**[1,3,*] **Shinichi Nakajima**[2,3,6,*]
[1]Fraunhofer HHI, [2]TU Berlin, [3]Berlin Big Data Center, [4]Korea University, [5]MPI for Informatics, [6]RIKEN AIP
{vignesh.srinivasan, wojciech.samek}@hhi.fraunhofer.de
{klaus-robert.mueller, nakajima}@tu-berlin.de

Abstract

Unpaired image-to-image domain translation involves the task of transferring an image in one domain to another domain without having pairs of data for supervision. Several methods have been proposed to address this task using Generative Adversarial Networks (GANs) and cycle consistency constraint enforcing the translated image to be mapped back to the original domain. This way, a Deep Neural Network (DNN) learns mapping such that the input training distribution transferred to the target domain matches the target training distribution. However, not all test images are expected to fall inside the data manifold in the input space where the DNN has learned to perform the mapping very well. Such images can have a poor mapping to the target domain. In this paper, we propose to perform Langevin dynamics, which makes a subtle change in the input space bringing them close to the data manifold, producing *benign examples*. The effect is significant improvement of the mapped image on the target domain. We also show that the score function estimation by denoising autoencoder (DAE), can practically be replaced with any autoencoding structure, which most image-to-image translation methods contain intrinsically due to the cycle consistency constraint. Thus, no additional training is required. We show advantages of our approach for several state-of-the-art image-to-image domain translation models. Quantitative evaluation shows that our proposed method leads to a substantial increase in the accuracy to the target label on multiple state-of-the-art image classifiers, while qualitative user study proves that our method better represents the target domain, achieving better human preference scores.

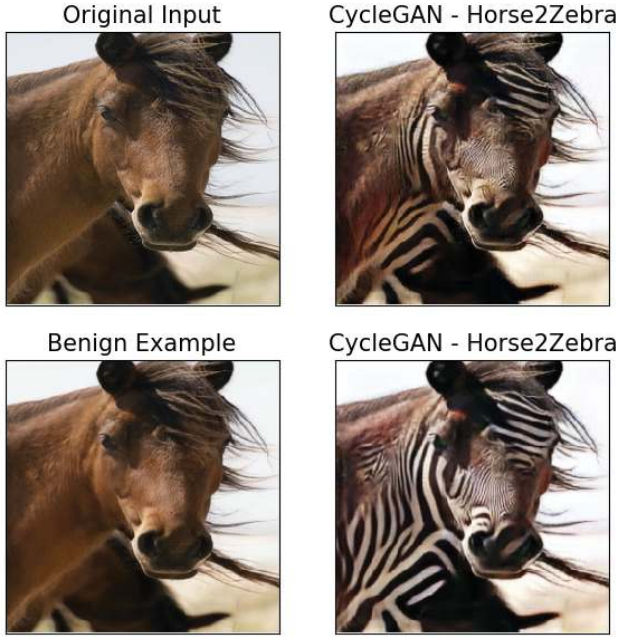


Figure 1: CycleGAN (Zhu et al. 2017) translates a horse image to a zebra image, which can however be unsuccessful if a test sample lies on the fringes of training distribution (top row). Our proposed ImproveIT makes an imperceptible change towards high density areas in the input space, leading to a significantly improved translated image (bottom row).

Introduction

Image translation (IT) is the problem of translating an image from one representation to another. Problems of this nature, like denoising, super-resolution, coloring, inpainting, style transfer among others are often ill-posed and require supervision for learning the mapping. Cross-domain IT is a more challenging task as it is very hard or close to impossible to find paired data.

Convolutional Neural Networks have achieved significant strides in addressing problems in IT. Several strategies were proposed to address the learning problem of IT, including Denoising Autoencoders (Vincent et al. 2008a) (Vincent et al. 2010), Variational Autoencoders (Kingma and Welling 2013), perceptual loss (Johnson, Alahi, and Fei-Fei 2016), U-nets (Ronneberger, Fischer, and Brox 2015) among others. However, the task of cross-domain IT still required paired data for the learning. Recent works propose to address this problem by unsupervised learning with the help of Generative Adversarial Networks (GAN). The task of translation is performed by an encoder. The generated images are then given to a discriminator as well as a decoder. The dis-

*Asterisks indicate corresponding author

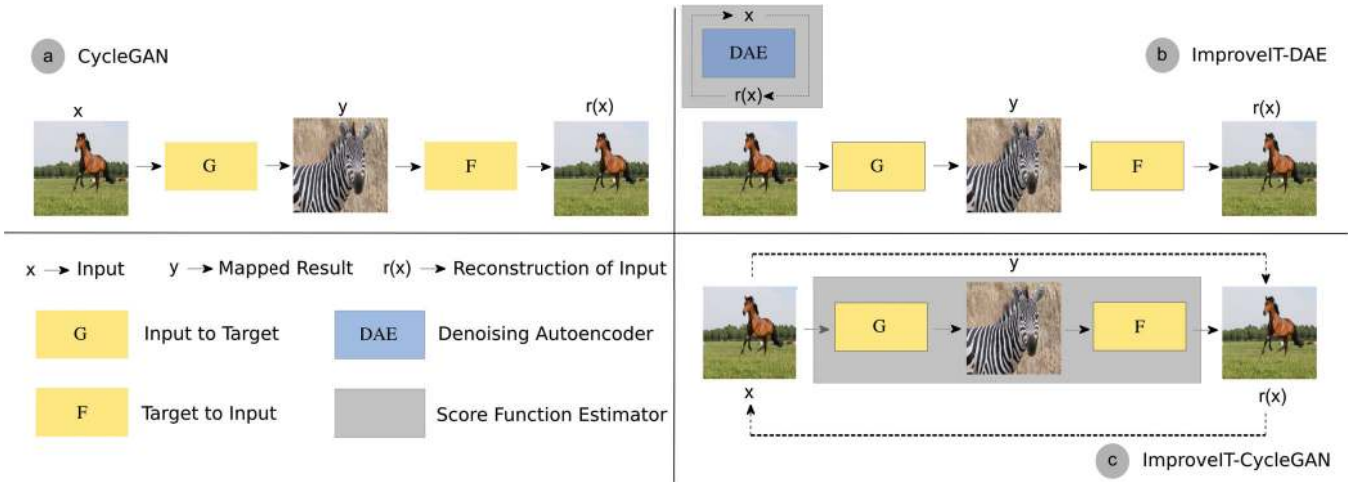


Figure 2: The different model assumptions that we considered. (a) The models used in CycleGAN (Zhu et al. 2017). (b) A DAE is trained separately on the training images of the horse dataset. This DAE is then used to estimate the score function during test time. (c) In essense, the cycle consistency constraint creates an autoencoding structure. $G : x \rightarrow y$ acts as an encoder while $F : y \rightarrow x$ acts as a decoder.

criminator performs the task of distinguishing the real images from the target domain to the fake images generated. The decoder on the other hand, projects back to the original domain. This mapping, also called as cycle consistency or reconstruction, ensures that the mapping to different domain retains the contents of the original image. CycleGAN (Zhu et al. 2017), (Kim et al. 2017) and (Yi et al. 2017) utilized cycle consistency constraint along with GANs to learn cross domain mapping. Although existing methods showed surprisingly successful examples, they do not always work – image translation typically fails on a significant proportion of the test samples.

In this paper, we hypothesize that the failures are due to the lack of training at the test point, and propose a method that reduces the proportion of unsuccessful test examples. Given a learned mapping from input domain to a target domain, we first detect test samples which lie on the fringes of the data distribution, which we call them as *fringe examples*. For the fringe examples, we perform Metropolis Adjusted Langevin Algorithm (MALA) with lower temperature to slightly nudge the fringe examples. This way we move the fringe examples to the high density regions of the data generating distributions, where the IT network is well trained on. With a small step size and the perturbation variance appropriately set, the samples are very similar to the original input. The perturbations are imperceptible to a human eye. Such a small change on the input domain, in turn brings about a drastic improvement on the target domain. We call such samples generated from the above mentioned MCMC sampler as *benign examples*. The images on the target domain generated from benign examples show improved features of target domain. Our method – ImproveIT (Improve Image Translation) – can be successfully used on top of other cross domain IT methods either by training a DAE separately or by utilizing the autoencoding structure already present in most IT methods. Using extensive quantitative experiments, we show that ImproveIT outperforms methods on which it is integrated into. Qualitative evaluations reveals consistent increase of the target domain attributes in the translated images.

Related Work

Pix2pix (Isola et al. 2017) utilized conditional GANs to effectively learn the task of IT. Concurrent works by (Zhu et al. 2017), (Kim et al. 2017), (Yi et al. 2017) propose to perform cross domain IT by removing the need for supervision (Isola et al. 2017), (Wang et al. 2018). The need for paired data was circumvented by using cycle-consistency loss which enforces that the image be translated back to the input domain. This constraint of reconstruction of the original image provides for the content of the image to be preserved in both the domains. (Liu, Breuel, and Kautz 2017) proposed sharing the latent space between the two domains, which implies cycle-consistency constraint. However, the inherently ill-posed problem of translating an image to another domain suffers from unimodality of the generated images. While (Zhu et al. 2017) translates a horse to a zebra (texture transformations), the method suffers to translate a cat to a dog (geometrical transformations). (Huang et al. 2018) and (Lee et al. 2018) propose splitting the latent space into two - a content space and a style space. The content space is shared between the two domains but the style space is unique to each domain. The style space is modelled with the assumption of a Gaussian prior. This helps in generating diverse images at test time by randomly sampling from a Gaussian distribution. Compressing the image into a latent space however makes the decoding process more challenging and affects the quality of the images generated. While (Huang et al. 2018) and (Kim et al. 2019) utilize adaptive instance normalization layer (AdaIN) to better model the style, (Kim et al. 2019) and (Mejjati et al. 2018) also makes use attention model. Inspite of the various techniques proposed to ad-

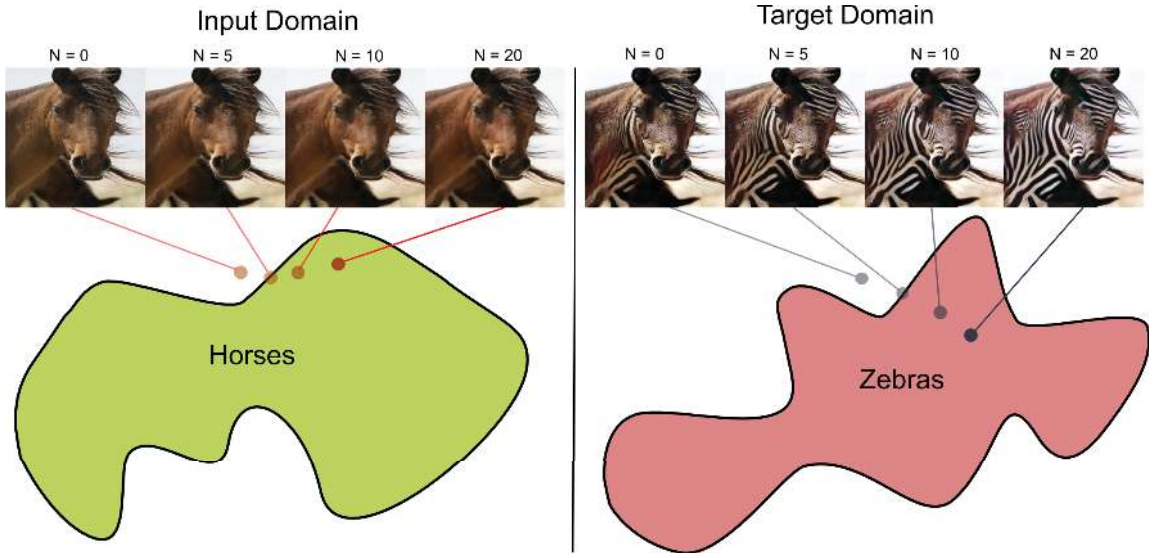


Figure 3: An intuitive visualization of the Langevin dynamics on the manifold of horses and its impact on the mapping to zebra space. As the test input (red dot) is sampled, it takes a step towards the manifold at every iteration.

dress the issues of cross domain IT, the unified framework of Pix2pix (Isola et al. 2017) and CycleGAN (Zhu et al. 2017) are still considered the state-of-the-art at cross domain IT for many tasks.

Background

In this section, we introduce two existing methods, on which our proposed method is based.

Metropolis-adjusted Langevin Algorithm

Metropolis-adjusted Langevin algorithm (MALA) is an efficient Markov chain Monte Carlo (MCMC) sampling method which uses the gradient of the energy (negative log-probability $E(\boldsymbol{x}) = -\log p(\boldsymbol{x})$). Sampling is performed sequentially by

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \alpha \boldsymbol{\nabla}_{\boldsymbol{x}} \log p(\boldsymbol{x}_t) + \boldsymbol{\nu}, \tag{1}$$

where α is the step size, and $\boldsymbol{\nu}$ is random perturbation subject to $\mathcal{N}(\mathbf{0}, \delta^2 \boldsymbol{I}_L)$. By appropriately controlling the step size α and the noise variance δ^2, the sequence is known to converge to the distribution $p(\boldsymbol{x})$.[1] (Nguyen et al. 2016) successfully generate realistic artificial images that follow the natural image distribution with the gradient estimated by denoising autoencoders.

Denoising Autoencoders

A denoising autoencoder (DAE) (Vincent et al. 2008b; Bengio et al. 2013) is trained so that data samples contaminated with artificial noise is cleaned. More specifically, it minimizes the reconstruction error:

$$(\boldsymbol{r}) = \mathbb{E}_{p'(\boldsymbol{x})p'(\boldsymbol{\nu})} \left[\|\boldsymbol{r}(\boldsymbol{x} + \boldsymbol{\nu}) - \boldsymbol{x}\|^2 \right] \tag{2}$$

where $\mathbb{E}_p[\cdot]$ denotes the expectation over the distribution p, $\boldsymbol{x} \in \mathbb{R}^L$ is a training sample subject to a distribution $p(\boldsymbol{x})$, and $\boldsymbol{\nu} \sim \mathcal{N}_L(\mathbf{0}, \sigma^2 \boldsymbol{I})$ is artificial Gaussian noise with mean zero and variance σ^2. $p'(\cdot)$ denotes an empirical (training) distribution of the distribution $p(\cdot)$, namely, $\mathbb{E}_{p'(\boldsymbol{x})}[g(\boldsymbol{x})] = N^{-1} \sum_{n=1}^{N} g(\boldsymbol{x}^{(n)})$ where $\{\boldsymbol{x}^{(n)}\}_{n=1}^{N}$ are the training samples. (Alain and Bengio 2014) discussed relation between DAEs and contractive autoencoders (CAEs), and proved the following useful property of DAEs:

Proposition 1 *(Alain and Bengio 2014) Under the assumption that* $\boldsymbol{r}(\boldsymbol{x}) = \boldsymbol{x} + o(1)$ *the minimizer of the DAE objective* (2) *satisfies*

$$\boldsymbol{r}(\boldsymbol{x}) - \boldsymbol{x} = \sigma^2 \boldsymbol{\nabla}_{\boldsymbol{x}} \log p(\boldsymbol{x}) + o(\sigma^2), \tag{3}$$

as $\sigma^2 \to 0$.

Proposition 1 states that a DAE trained with a small σ^2 can be used to estimate the gradient of the log probability.

(Nguyen et al. 2016) utilized MALA to explore the latent space of a generator to produce more diverse images. The estimation of the score function was made possible by considering the generator of a GAN architecture as the encoder and a pretrained classifier (CaffeNet) as the decoder of the DAE. Although, this was not a formal DAE, the assumption was used successfully to explore the latent space. (Srinivasan et al. 2018), on the other hand utilized MALA on the input image space to a classifier for the purpose of defense against adversarial examples. However, in this case, a supervised DAE was used to estimate the score function of the joint distribution instead of the marginal distribution.

[1] For convergence, a rejection step after Eq (1) is required. However, it was observed that a variant, called MALA-approx (Nguyen et al. 2016), without the rejection step gives reasonable sequence for moderate step sizes. We use MALA-approx in our proposed method.

Proposed Method

In this section, we explain how we use the two existing methods, i.e., Metropolis-adjusted Langevin algorithm and denoising autoencoders, and then propose our method, called ImproveIT, for image translation.

Cooling Down by Langevin Dynamics

Our idea is to drive *fringe* samples towards the high density area, in which the network is expected to be trained well. This can be achieved simply by applying MALA (1) to each sample, with the step size α and the variance of the random perturbation satisfying

$$\alpha > \delta^2/2. \tag{4}$$

Since we use MALA without rejection step, i.e., MALA-approx, it is a discrete approximation to the (continuous) Langevin dynamics:

$$\frac{d\boldsymbol{x}}{dt} = \boldsymbol{\nabla}_{\boldsymbol{x}} \log p(\boldsymbol{x}) + \sqrt{2}\frac{d\boldsymbol{W}}{dt}, \tag{5}$$

if $\delta = \sqrt{2\alpha}$, where $\boldsymbol{W}$ is the Brownian motion. The dynamics (5) is known to converge to $p(\boldsymbol{x})$ as the equilibrium distribution (Roberts and Tweedie 1996; Roberts and Rosenthal 1998). By setting the step size and the perturbation variance so that Inequality (4) holds, we can approximately draw samples from the distribution with *lower temperature*, as shown below.

By seeing the negative log probability as the energy $E(\boldsymbol{x}) = -\log p(\boldsymbol{x})$, we can see $p(\boldsymbol{x})$ as the Boltzmann distribution with the inverse temperature equal to $\beta = 1$:

$$p_\beta(\boldsymbol{x}) = \frac{1}{Z_\beta} \exp\left(-\beta E(\boldsymbol{x})\right), \tag{6}$$

where $Z_\beta = \int \exp\left(-\beta E(\boldsymbol{x})\right) d\boldsymbol{x}$ is the partition function. We have the following theorem:

Theorem 1 *In the limit when $\alpha, \delta^2 \to 0$ with their ratio α/δ^2 kept constant, the sequence of the MALA-approx* (1) *converges to $p_\beta(\boldsymbol{x})$ for*

$$\beta = \frac{2\alpha}{\delta^2}. \tag{7}$$

(Proof) As α and δ^2 go to 0, the MALA-approx (1) converges to the following dynamics:

$$\frac{d\boldsymbol{x}}{dt} = \boldsymbol{\nabla}_{\boldsymbol{x}} \log p(\boldsymbol{x}) + \frac{\delta}{\sqrt{\alpha}}\frac{d\boldsymbol{W}}{dt},$$

which is equivalent to

$$\frac{d\boldsymbol{x}}{dt} = \frac{2\alpha}{\delta^2}\boldsymbol{\nabla}_{\boldsymbol{x}} \log p(\boldsymbol{x}) + \sqrt{2}\frac{d\boldsymbol{W}}{dt}. \tag{8}$$

Eq.(8) can be rewritten with the Boltzmann distribution (6) with the inverse temperature specified by Eq.(7):

$$\frac{d\boldsymbol{x}}{dt} = \boldsymbol{\nabla}_{\boldsymbol{x}} \log p_\beta(\boldsymbol{x}) + \sqrt{2}\frac{d\boldsymbol{W}}{dt}.$$

Comparing it with Eq.(5), we find that this dynamics converges to the equilibrium distribution $p_\beta(\boldsymbol{x})$, which completes the proof. □

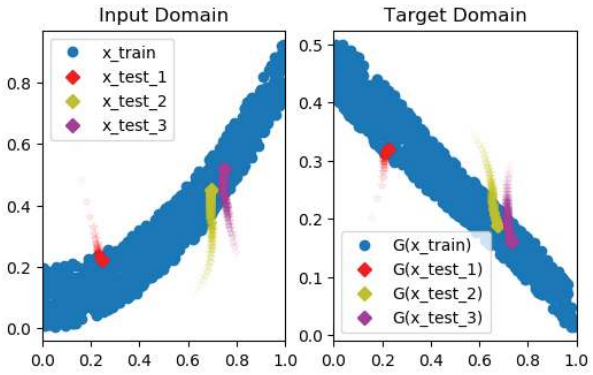


Figure 4: Simulation of CycleGAN (Zhu et al. 2017) on a two dimensional toy dataset. The input is mapped from one domain to another in an unsupervised setting using G, a two layer feed forward network, with the same loss functions as used in (Zhu et al. 2017). Three test points are considered from outside of the training data manifold. A visualization of the Langevin dynamics on these test samples is shown here.

Theorem 1 states that the ratio between α and δ^2 effectively controls the temperature. Specifically, we can see the MALA-approx (1) as a discrete approximation to the Langevin dynamics converging to the distribution given by

$$p_{2\alpha/\delta^2}(\boldsymbol{x}) = \frac{p^{2\alpha/\delta^2}(\boldsymbol{x})}{\int p^{2\alpha/\delta^2}(\boldsymbol{x})d\boldsymbol{x}},$$

of which the probability mass is more concentrated than $p(\boldsymbol{x})$ if Inequality (4) holds. This way, we use the MALA to drive test samples towards high density area, before performing the image translation.

Fringe Examples Detection

Test samples come from relatively low density regions of the input distribution. Such fringe examples can be mapped poorly on the target domain by learned mapping. However, not all test samples will be off the manifold. Applying Langevin dynamics to samples already on the manifold might not be beneficial in improving the mapping to the target domain. Hence, we propose to detect if the test samples lie on the fringes of the data manifold. Given a test sample, we estimate the score function, which gives information about the distance of the test sample from the manifold.

$$\|\boldsymbol{\nabla}_{\boldsymbol{x}} \log p(\boldsymbol{x})\| > \varepsilon \tag{9}$$

If the score function is above a predetermined threshold, then we consider them as fringe examples and perform Langevin dynamics to move them towards the underlying data distribution. If not, we consider that the mapping is already ideal and leave the sample as it is.

ImproveIT

In our proposed method, called ImproveIT, we apply fringe sample detection and Langevin dynamics cooling before ap-

Table 1: Classification accuracy in percentage on the fake zebras generated from CycleGAN and the accuracy of zebras generated from ImproveIT are shown here. The suffix to ImproveIT corresponds to the number of steps of the random walk. Every step of the random walk helps in making the mapping to the target zebra domain more successful. Hence, the classification accuracy on the fake zebras increases as the steps increases. Increasing the number of steps can lead to further increase in the accuracy. However, the change in the original space (compared to the original) gets perceptible because Langevin dynamics can bring samples to anywhere within the cluster after many steps. It is to be noted that none of the classifiers shown here are used in the training of CycleGAN.

Classifier	CycleGAN	ImproveIT-20	ImproveIT-40	ImproveIT-60	ImproveIT-80	ImproveIT-100
VGG16	85.00	87.50	88.33	89.16	90.83	**92.42**
InceptionV3	85.00	85.83	86.67	87.50	88.33	**93.93**
ResNet18	82.50	85.00	85.83	86.67	87.50	**92.42**
ResNet50	89.16	90.00	90.83	90.83	91.66	**93.93**
ResNet101	89.16	89.16	90.00	90.00	90.83	**92.42**
DenseNet169	88.33	90.00	92.50	92.50	91.66	**93.93**
DenseNet201	87.50	90.00	90.00	90.00	90.00	**93.93**

plying image translation. Namely, we first apply the thresholding on the norm of score function. If the test sample satisfies (9), we apply MALA-approx with the step size and the perturbation variance satisfying Inequality (4). The resulting sample is fed into the image translation network.

To perform ImproveIT, the score function can be estimated by training a DAE on the set of training images of input domain. We call this variant as ImproveIT-DAE. However, most IT methods intrinsically have an autoencoding structure due to the cycle consistency constraint. The translation to the target domain by G forms the encoder, while the translation back to the input domain forms the decoder. Motivated by (Nguyen et al. 2016), where the score function is estimated by the autoencoder consisting of a GAN architecture as the encoder and a pretrained classifier (CaffeNet) as the decoder of the DAE, we propose another variant, called ImproveIT-CycleGAN. Namely, we estimate the score function by the reconstruction residual of CycleGAN, which removes the necessity of training an additional DAE.

Analysis

Toy Data

To demonstrate the concept of our proposed method ImproveIT, we define a toy setting where samples from one domain are mapped to another domain. A two-layer feed forward network is used to perform the mapping. The loss functions used are exactly the same as used in (Zhu et al. 2017) to simulate the same environment. After training the network, three points outside the input data manifold are considered for the test. All three are mapped to the target domain as shown in the Figure 4. We find that the mapping also lies outside of the target data manifold. The network is well trained only on high density region of the data manifold (shown in blue in Figure 4). Now, we perform Langevin dynamics with these three points on the input domain as the initial points. We find that the samples are driven to the high density region of the data manifold on the input domain. Mapping these samples through the trained feed forward network shows us that the mapping on the target domain is now also driven towards the high density region of the target data manifold.

This forms the crux of our proposal that once a model is well trained, it can happen that, at test time an image is provided, which lies on the fringes of the training data distribution. This can have the effect that the input is mapped poorly to the target domain. We find that moving the input towards the data distribution on the input domain can have the effect of improving the mapping in the target domain.

(a) CycleGAN (b) ImproveIT-DAE (c) ImproveIT-CycleGAN

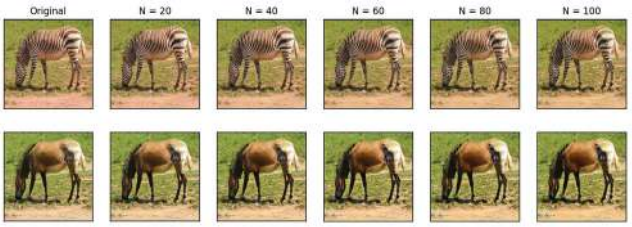
Figure 5: A separately trained DAE can be used as score function estimator. On the other hand, $G + F$ in CycleGAN can also be used to do the same. Increase in the target domain attribute on the resulting images of ImproveIT-DAE and ImproveIT-CycleGAN can be seen here.

ImproveIT-DAE $\rightarrow$ ImproveIT-CycleGAN

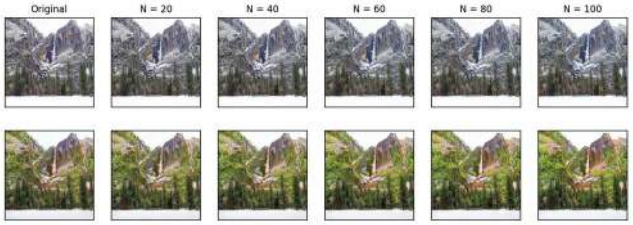
As first experiment, we reproduce the results of CycleGAN (Zhu et al. 2017) by using their pretrained models available publicly on Github[2]. We apply Langevin dynamics on the input space by incorporating ImproveIT into CycleGAN. The estimation of the score function can be performed by using a DAE trained separately. However, in most IT methods, there exist the cycle consistency constraint. The input image is mapped to the target domain by $G : \boldsymbol{x} \rightarrow \boldsymbol{y}$. And the input is reconstructed from the mapping by $F : \boldsymbol{y} \rightarrow \boldsymbol{x}$. Thus, $G + F$ can be combined to form an autoencoder. Note that this will however not be a theoretically formal DAE. The models were trained with GAN loss along with L_1 distance

[2]https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix/

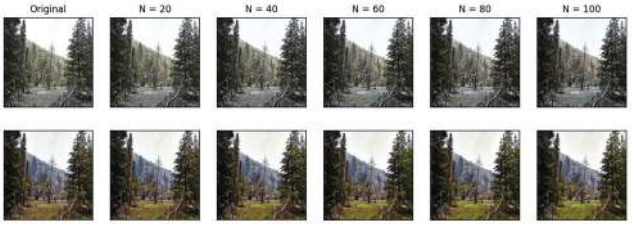
(a) Horse2Zebra: Expanding of zebra stripes over the horse.



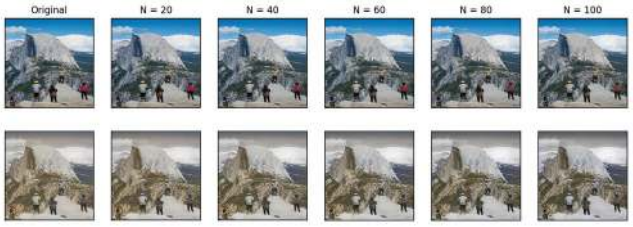
(b) Zebra2Horse: The mapped horse becomes slightly more brown than the result of CycleGAN



(c) Winter2Summer: The image gets enhanced with greener patches at many places.

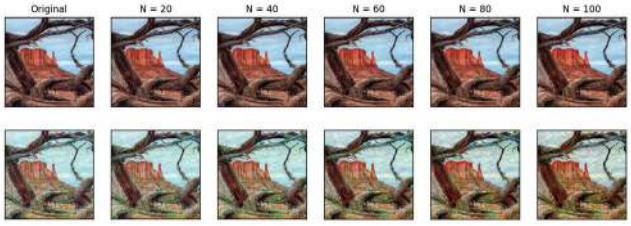


(d) Winter2Summer: The grass becoming greener along the image gets brighter.

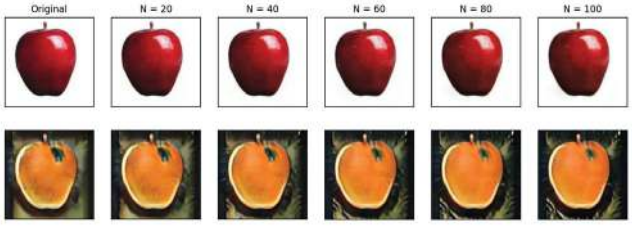


(e) Summer2Winter: The sky gets darker along with the mountain

(f) Summer2Winter: Sky and clouds gets darker.



(g) Style2Monet: With more steps, the sky is filled with brush stroke like patterns to make it more like a painting.



(h) Apple2Orange: The artefacts are reduced with an increase in orange color.

Figure 6: Example results of applying ImproveIT on the models of CycleGAN (Zhu et al. 2017) for different tasks. In each image, the top row represents the sampling on the input domain. Each sampled image is then mapped to the target domain using the pretrained model. The results of the mapping are shown on the bottom row. The iteration number is mentioned in the top row above each image. ImproveIT brings very little change to the input while showing enhancement of the target domain attributes in the resulting images.

for the reconstruction measure which is contradictory to the L_2 distance used for training a DAE. However, we ascertain that any autoencoding structure which minimizes the reconstruction error, can provide an approximate path towards the high density region of the data distribution.

Our experiments on the Horse2Zebra task of CycleGAN reveal this. We perform Langevin dynamics on the test inputs of horses using both the methods and obtain the final result of the mapping on the target domain. In both the cases, a step size of $\alpha = 0.01$ and standard deviation of $\sigma = 0.001$ with a fixed number of iterations $N = 50$ was used for the Langevin dynamics.

As shown in Figure 5, both ImproveIT-DAE and ImproveIT-CycleGAN increase the target domain attribute

over CycleGAN in the output image, which is "Zebra" here. However, ImproveIT-CycleGAN has the innate advantage that it needs no additional training. In cases, where training data is not available, the autoencoding structure can be used freely. Henceforth, in all our remaining experiments we only consider ImproveIT-CycleGAN, unless specified otherwise.

Experiments

Quantitative Evaluation

Accuracy on Pretrained Image Classifiers To show that ImproveIT improves the results on the target domain by moving the mapping towards the target data distribution, we perform quantitative evaluation by passing the generated images from the entire test set through state-of-the-art pretrained image classifiers. All the classifiers – VGG16 (Simonyan and Zisserman 2014), Resnet18 (He et al. 2016), InceptionV3 (Szegedy et al. 2016) and Resnet50 (Xie et al. 2017), Resnet101 (Zagoruyko and Komodakis 2016), DenseNet169 and DenseNet201 (Huang et al. 2017) have been trained on the Imagenet dataset (Deng et al. 2009) which includes a label for "Zebras".

The resulting Top-1 accuracy on the fake zebra images from CycleGAN is shown in the second column in Table 1. The result of using ImproveIT with varying number of iterations is also shown. The fake zebras of CycleGAN obtain a classification accuracy of around 86%. ImproveIT performs the random walk on the input space and these samples mapped to the zebra domain consistently improve the classification accuracy on every classifier reaching 93% on average over the different classifiers. The result reiterates our proposal that perturbing the input with ImproveIT drives the mapping closer to the target data distribution.

Sat2Map In order to make a quantitative evaluation on paired data at test time, we used the map dataset (Isola et al. 2017) consisting of satellite images and corresponding maps. The translation from satellite images to maps is unimodal and hence, we evaluate the performance of ImproveIT in this setting. We consider the pretrained models of Pix2pix (Isola et al. 2017) and CycleGAN (Zhu et al. 2017) available publicly on Github. In this evaluation, we make use of the fringe examples detection strategy. Langevin dynamics is performed only on those test samples whose score function is higher than a given threshold. This prevents applying the sampling method on those test samples which are already on the manifold. The satellite images are translated into maps and these images are compared with the ground truth maps. A pixel was counted as correct if the distance between the color values of the generated image and the ground truth was below 16, similar to the evaluation strategy in (Liu, Breuel, and Kautz 2017).

In Table 2, the mean pixel-wise accuracy is shown for the results of Pix2pix, CycleGAN and for ImproveIT which is utilized over each method. ImproveIT bring about small, yet significant improvement in the accuracy of the maps. Figure 7 shows the output of CycleGAN followed by the enhancement provided by ImproveIT in comparison with the ground truth.

(a) CycleGAN (b) ImproveIT (c) GT

Figure 7: The result of CycleGAN and of using ImproveIT is shown here for the task of Sat2Map in comparison with the ground truth (GT). After fringle example detection, the input to the above image was sampled with Langevin dynamics. The result is the enhancement on the target domain, which can be visually noticed when comparing it to the GT.

Table 2: Accuracy of translation from satellite images to maps using 2 state-of-the-art image translation methods – Pix2pix (Isola et al. 2017) and CycleGAN (Zhu et al. 2017) – is shown here. ImproveIT is utilized for each of the methods and the corresponding accuracies are also shown.

Model	Accuracy
Pix2pix	58.04
ImproveIT-Pix2pix	**58.38**
CycleGAN	70.68
ImproveIT-CycleGAN	**71.06**

Qualitative Evaluation

User Study To quantitatively evaluate the images generated as a result of using ImproveIT, we performed a user study on Amazon Mechanical Turk (AMT) for human preference. Users were asked to pick either of the two images as to which one represented most the target domain – results of CycleGAN or the results of ImproveIT. We performed these experiments on two different tasks of CycleGAN as shown in Table 3. Each experiment involved 25 users who were shown 50 images in total. The first 10 images were for practice and the remaining 40 images were considered for computing the preference score. The feedback from the practice session decided whether to consider the participant's preferences in the final score (Zhu et al. 2017). A preference score of 50% indicates that ImproveIT does not provide any additional benefits to the result of CycleGAN. Below 50% means that ImproveIT deteriorates the results of CycleGAN mapping. As shown in Table 3, users preferred that the images generated from ImproveIT represented the target domain more than CycleGAN.

It can also be visually seen in Figure 6 for the various tasks. In the case of Horse2Zebra, the poor mapping of CycleGAN is improved at every step of the Langevin dynamics by ImproveIT by increasing the zebra strips as shown in Figure 6a. The Zebras in Figure 6b are mapped iteratively to a horse which becomes more brown. For Winter2Summer, the final output image at $N = 100$ in Figure 6c and Figure 6d show clear increase in the greenery (indicating more sum-

Table 3: User Study on AMT for human preference scores. Users were asked to pick one of two images – results of CycleGAN or result of ImproveIT for the tasks of Horse2Zebra and Winter2Summer. For Horse2Zebra task, we also perform denoising study. The input here is distorted artificially by adding salt and pepper noise to some percentage of the pixels. The fake zebras generated from the distorted inputs for CycleGAN and ImproveIT were then posted on AMT for human preference.

Task	CycleGAN	ImproveIT
Horse2Zebra	34	**66**
Winter2Summer	37	**63**
Denoising Study		
% of pixels corrupted	CycleGAN	ImproveIT
5	27	**73**
10	20	**80**
25	27	**73**
20	21	**79**

mer) compared to the result of CycleGAN. This is in contradiction to Summer2Winter, where the output image in the second row shows increase in grey shade with the sky getting darker at every step as shown in Figure 6e and Figure 6f. For the task of Style2Monet, Langevin dynamics beings about brush stroke patterns on the sky as shown in Figure 6g. In Figure 6h, the apple is translated to an orange, but the mapping by CycleGAN produces artefacts. Improve-IT reduces the artefacts while increasing the orange color in the target domain.

Denoising Study During inference, it cannot always be expected that the test image would lie on the training data distribution where the network has been well trained on. Hence, we simulate such an environment by artificially adding salt and pepper noise to the input images for the task of Horse2Zebra. The mapping of CycleGAN becomes very ineffective. Applying ImproveIT on the input space, then performs the task denoising on the input image. Denoising is a process which implicitly dives the sample towards the data distribution. Hence, the resulting fake zebras obtained by using ImproveIT are mapped better to the target domain. To evaluate the results, we performed a study with 25 users on AMT for human preference. Users were asked to pick either of the two images – results of CycleGAN or the results of applying ImproveIT. As shown in Table 3, users consistently preferred the results of ImproveIT over the results of CycleGAN.

Conclusion

Image translation is a challenging problem, especially for mapping from one domain to another. Several methods proposed include a cycle consistency constraint to learn in an unsupervised manner. We presented a general framework to make any learned translation model give better mappings on the target domain at test time. We propose to (a) identify *fringe examples*, that lie just outside of the data manifold and (b) bring them towards the manifold by using Langevin dynamics. This leads to the creation of *benign examples* which look very similar to the original input but give significantly better results when mapped to the target domain. We show that score function approximation can be obtained without training any additional models by using the cycle consistency architecture of the image translation models. Benign examples show the opposite effect of adversarial examples and may prove beneficial to many existing algorithms. It can also be another evaluation strategy for new algorithms to improve their results.

Acknowledgements

The authors thank Dr. Pan Kessel for discussion on Langevin dynamics. This work was supported by the German Ministry for Education and Research as Berlin Big Data Center (01IS18025A) and the Berlin Center for Machine Learning (01IS18037I), and by the Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (No. 2017-0-00451).

References

Alain, G., and Bengio, Y. 2014. What regularized auto-encoders learn from the data-generating distribution. *The Journal of Machine Learning Research* 15(1):3563–3593.

Bengio, Y.; Yao, L.; Alain, G.; and Vincent, P. 2013. Generalized denoising auto-encoders as generative models. In *Advances in Neural Information Processing Systems*, 899–907.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708.

Huang, X.; Liu, M.-Y.; Belongie, S.; and Kautz, J. 2018. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 172–189.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, 694–711. Springer.

Kim, T.; Cha, M.; Kim, H.; Lee, J. K.; and Kim, J. 2017. Learning to discover cross-domain relations with generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 1857–1865. JMLR. org.

Kim, J.; Kim, M.; Kang, H.; and Lee, K. 2019. U-GAT-IT: unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. *CoRR* abs/1907.10830.

Kingma, D. P., and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Lee, H.-Y.; Tseng, H.-Y.; Huang, J.-B.; Singh, M.; and Yang, M.-H. 2018. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 35–51.

Liu, M.-Y.; Breuel, T.; and Kautz, J. 2017. Unsupervised image-to-image translation networks. In *Advances in neural information processing systems*, 700–708.

Mejjati, Y. A.; Richardt, C.; Tompkin, J.; Cosker, D.; and Kim, K. I. 2018. Unsupervised attention-guided image-to-image translation. In *Advances in Neural Information Processing Systems*, 3693–3703.

Nguyen, A.; Yosinski, J.; Bengio, Y.; Dosovitskiy, A.; and Clune, J. 2016. Plug & play generative networks: Conditional iterative generation of images in latent space. *arXiv preprint arXiv:1612.00005*.

Roberts, G. O., and Rosenthal, J. S. 1998. Optimal scaling of discrete approximations to Langevin diffusions. *Journal of the Royal Statistical Society, Series B* 60:255–268.

Roberts, G. O., and Tweedie, R. L. 1996. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli* 2:341–363.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Srinivasan, V.; Marban, A.; Müller, K.-R.; Samek, W.; and Nakajima, S. 2018. Robustifying models against adversarial attacks by langevin dynamics. *arXiv preprint arXiv:1805.12017*.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

Vincent, P.; Larochelle, H.; Bengio, Y.; and Manzagol, P.-A. 2008a. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, 1096–1103. ACM.

Vincent, P.; Larochelle, H.; Bengio, Y.; and Manzagol, P.-A. 2008b. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, 1096–1103. ACM.

Vincent, P.; Larochelle, H.; Lajoie, I.; Bengio, Y.; and Manzagol, P.-A. 2010. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research* 11(Dec):3371–3408.

Wang, T.-C.; Liu, M.-Y.; Zhu, J.-Y.; Tao, A.; Kautz, J.; and Catanzaro, B. 2018. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8798–8807.

Xie, S.; Girshick, R.; Dollár, P.; Tu, Z.; and He, K. 2017. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1492–1500.

Yi, Z.; Zhang, H.; Tan, P.; and Gong, M. 2017. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE international conference on computer vision*, 2849–2857.

Zagoruyko, S., and Komodakis, N. 2016. Wide residual networks. *arXiv preprint arXiv:1605.07146*.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232.