This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Unsupervised Learning of Paired Style Statistics for Unpaired Image Translation

Saeid Motiian[1] Quinn Jones[2] Stanislav Pidhorskyi[2] Gianfranco Doretto[2]
[1] Adobe Applied Research [2] West Virginia University
motiian@adobe.com {qjones1, stpidhorskyi, gidoretto}@mix.wvu.edu

Abstract

Image-to-image translation has the goal of learning how to transform an input image from one domain as if it was from another domain, while preserving semantic and global information from the input. We present an image-to-image translation method that can be trained with unpaired images from source and target domains. However, we introduce a regularization that allows the model to specifically translate the local spatial statistic from one domain to another in an effort to leave unchanged gross structures and discourage translations of the semantic content. We do so by learning to generate paired images mapping the local statistic from one domain to the other. In turn, such images are used to improve the training of the translation networks, which become more focused on translating only the "style" of images while preserving the semantic content. Experiments on domain translation as well as domain adaptation highlight the effectiveness of our approach in comparison with the state-of-the-art.

1. Introduction

We present a new method for unsupervised image-to-image translation using a Generative Adversarial Network (GAN) [11] based method. Unpaired image translation seeks to learn how to take an image from one domain (e.g., images of horses) and reproduce it as if it was from another (e.g., images of zebras). This can be applied to many different computer vision problems, such as domain adaptation [3] or image modification, like changing a celebrity's facial expression [5], or super-resolution [18].

Unsupervised image-to-image translation usually suffers from two important problems. First, in absence of paired images in training, it is hard to teach the translation networks which are the parts of the scene that need to be translated. Cycle-consistency [43] based approaches exhibit that problem because they have no additional constraints addressing it, which is why some recent methods [29, 4] proposed to leverage attention mechanisms to teach networks where to attend.

While there are many approaches to image-to-image translation [41, 17, 24, 37, 44, 25], they do not directly address a second important issue, noted in [44, 15]. In particular, despite creating realistic target images, they may change their semantic content as a side effect of the translation process. For example, in a translation task between MNIST [21] and SVHN [32] datasets an image from label 3 may be used as source and is translated to SVHN and back and still resemble a 3, but in the SVHN domain it's semantic content will often flip to some other digit. This makes previous work unusable for mapping styles for domain adaptation [35, 20, 12, 10, 7, 38, 36], and could also generate issues with changing too much about an image in other tasks such as altering the outlines of a subject.

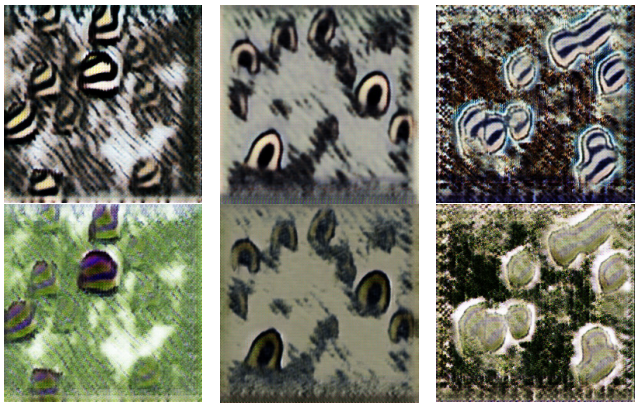
Figure 1: **Paired style image translation.** The first column shows a style image pair generated by our method. The second and third columns show examples of our results translating the horses (top-middle) to zebras (top-right) and vice versa. Note how the watermark in the top-middle column is unaffected as it is not in the style statistic for either domain.

In this work, we took inspiration from works in style transfer noting that, for instance, in translating an apple image to the orange's domain the expected result is to have the apples in the image more or less "replaced" by oranges with little else modified. This is why we propose to learn the joint statistics between the translation domains of the local "styles," and use it as a way to restrict the translation to color and texture modifications, rather than altering

gross structures. In absence of paired samples defining the domain local styles, we learn how to generate *paired style images*, together with the translation networks. The result is a regularized cycle-consistency model that can learn the correspondence between the textures of the two domains that need to be translated, versus those that should be maintained. The experiments show this makes for a translation module which preserves more semantic information, and gives us a model that is more adept at domain adaptation, while giving good results for other translation tasks.

2. Related Work

Image-to-Image Translation. Early formulations of the image-to-image translation problem can be traced back to [14], where they used handcrafted metrics to match sliding windows of texture patches from the source texture to the target image. However, this method requires having semantic pairs of samples to work with which can be more costly to setup compared to an unsupervised approach. To leverage unsupervised samples Generative Adversarial Networks (GAN) [11] have been the focus of most recent works, including this one. The adversarial training creates a scenario in which the generator can learn the distribution of realistic images and even semantic information about the images without the use of pairs. For instance, in ConditionalGAN [17] they train both the generator and the discriminator to have access to the source sample thus allowing the discriminator to evaluate the conditional probability of the generated sample. In [37] they showed that by constraining the generator to act as an identity function on their target domain they were able to translate portrait images of celebrities to caricatures of their faces, despite not having any correspondence in the training set. In [24] instead, they used a generator model based on Variational Auto-Encoders (VAE), which assumes that the unpaired images can both be compressed to the same latent code, while still being reconstructable from the decoder. And translational GANs have even been shown to augment data effectively enough to produce competitive results in domain adaptation tasks [3]. More recently, cycle-consistency [43] has become a popular approach for unpaired image-to-image translation. In the unsupervised setting, translation networks must additionally learn which parts of the scene are intended to be translated, when this is not directly enforced by the cycle-consistency. [29, 4] have applied attention mechanisms to each translation network, trying to enforce the networks to look at the desired parts of the scene. Our approach relates to those because it aims at achieving a similar effect, but we do not need an additional attention network for translation, and we regularize the cycle-consistency based on generating image pairs capturing the respective "styles" to be translated. Finally, [23] uses adversarial training with domain-specific information to perform continuous cross-domain image translation and manipulation.



(a) Zebra-Horse style pairs

(b) Apple-Orange style pairs

Figure 2: **Paired style images.** Paired style images from Zebra to Horse dataset (a), and Apple to Orange dataset (b). The top row shows examples of generated style images from the source domain and the second row shows the corresponding generated style images from the target domain.

Texture Style Transfer. Image-to-image translation and texture style transfer have been linked since the early works [14, 6] in that they share the ill-posed problem of preserving the overall structure of the input while modifying the colors, textures, and other attributes to create the desired effect on the output. One thing that was always clear is the importance of finding texture correspondences between the source image and the source texture at multiple scales. These ideas have evolved also onto more recent works leveraging deep networks [9], which use the Gram matrix to measure the correlation of textures at multiple scales within the image to capture and disentangle the image's "style". Subsequently, [28] took into consideration spatial correspondence through the use of semantic segmentation to further ensure that outputs are, for instance, photorealistic as opposed to picturesque. [40] instead, introduced an unsupervised learning approach to discover, summarize, and manipulate artistic styles from large collections of paintings. Differently than previous work, we are more interested in

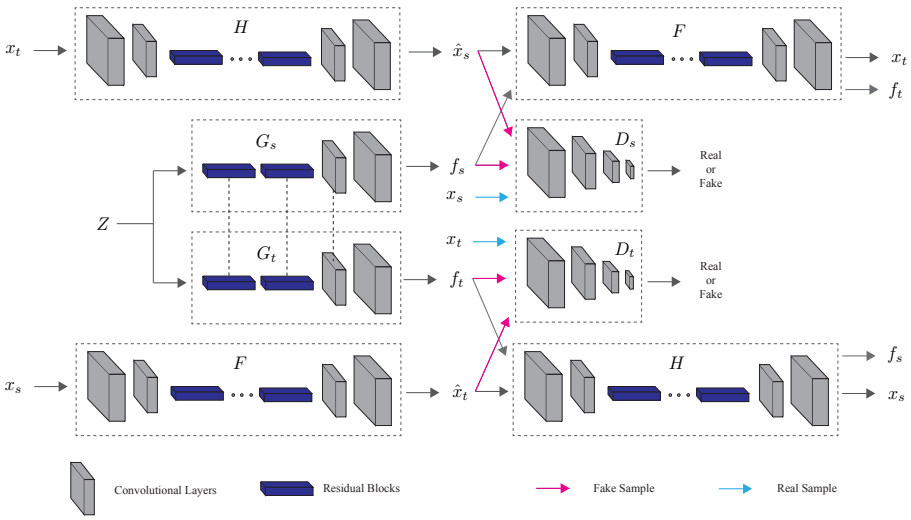


Figure 3: **End-to-end training architecture.** This plot shows the end-to-end training of our model. G_s and G_t generate the paired images f_s and f_t, respectively. Since f_s is paired with f_t, if we pass f_s to the mapping network F, we should be able to reconstruct f_t. Similarly, since f_t is paired with f_s, if we pass f_t to the mapping network H, we should be able to reconstruct f_s. In this way, F and H are trained directly and they are encouraged to correctly map the local statistics from source to target and vice versa while preserving the image semantic content during the translation. Most layers of G_s and G_t are shared in order to generate meaningful pairs. Each discriminator receives three sets of images. The first set includes real images of the source or target distribution (blue arrows). The second and third includes the translated images and generated images which make up the fake group (red arrows).

directly finding texture-texture correspondences which are present in the images of a source and a target domain, i.e., the texture of an apple skin (source) should become like the texture of an orange peel (target) in particular translation tasks, but the background textures while present should be unchanged as much as possible. This we do by learning the joint distribution of textures across the image domains.

3. Overview

Our focus is the image-to-image translation problem, where we are given a training dataset made of sets of images $\mathcal{D}_s = \{(x_i^s)\}_{i=1}^N$ and $\mathcal{D}_t = \{(x_i^t)\}_{i=1}^M$ referred to as the *source* and *target* datasets, respectively, and each x_i^s, x_i^t are realizations of random variables X^s, X^t, defined by the source and target distributions. Our goal is to translate an image drawn from the source distribution into an image that appears as if it was drawn from the target distribution.

This problem can be solved by training a conditional GAN architecture with pairs of corresponding images from the two domains [16]. However, collecting pairs can be costly or unrealistic in many situations. Therefore, several approaches have proposed to address the image-to-image translation problem in *unsupervised* settings, whereby no two training samples from source and target domains can be paired [16, 43, 24]. Relevant regularization strategies for addressing the absence of pairing information include imposing cycle-consistency [43], or weight-sharing [24].

Our approach aims at further alleviating the missing pairing information by adding to the training procedure paired images from source and target domains. Doing so should make the training more similar to learning a conditional GAN with paired images. However, since those pairs are not directly included in the training datasets, we propose to learn how to generate them. Models such as [25] can effectively generate such image pairs through the use of weight-sharing among generators and discriminators, but in order to generate image pairs of suitable resolution the generator networks must work in multiple stages [42, 19], extrapolating more image information in each layer. On the other hand, since we want to improve the translation, we do not necessarily need the generation to look realistic at every scale. This is why we leverage a simplified generator model which creates what we call *paired style images*. These are stochastically generated paired images that capture how the local spatial statistics of the source domain should be translated into the target domain, but do not represent a semantic translation at larger scales. Since the translation pipeline is supposed to be effective also on paired style images, the resulting model translates source images more reliably by replacing the local image statistics with those from the target domain.

For example, let us consider the task of translating images of zebras into images of horses, or vice versa. This could be done by finding the striped texture in the source image, which represents the zebras, and convert it to a uni-

(a) Horse to zebra. (b) Zebra to horse.

Figure 4: **Horse to zebra and zebra to horse translation.** Examples showing how a source image with horses is translated into a target image with zebra style (a), and vice versa (b). For every translation pair the left image is the source image, while the right image is the translated image.

form texture, which represents horses. Our model learns paired style generators capable of synthesizing paired style images as in Figure 2a. The top row shows generated style images of zebras and the bottom row their paired style images of horses. Each zebra style image contains several striped texture areas while its paired horse style image replaces the corresponding areas with more uniform texture in mottled brown. By training the translation network using these paired style images, we teach the network to look for and replace mainly the image "style" identified by the mapping between the local statistics of the two domains that has been learned by the generators, and is "visualized" as in Figure 2a. Figure 2b shows some generated paired style images for the task of translating images of apples into images of oranges.

4. Model Definition

Similar to [43], our model consists of two mapping networks $F : X^s \rightarrow X^t$ and $H : X^t \rightarrow X^s$. Our model also includes two generators $G_s : Z \rightarrow X^s$ and $G_t : Z \rightarrow X^t$ that generate paired style images (f_s, f_t) by randomly drawing from Z, a normally distributed variable. G_s and G_t share some weights as we will explain later. We also use two adversarial discriminators D_s and D_t. D_s aims at distinguishing local style statistics of source real images, x_s, from local style statistics of generated (fake) source images, f_s, and of translated images, $H(x^t)$. D_t instead, aims at distinguishing local style statistics of target images, x_t, from local style statistics of generated (fake) target images, f_t, and of translated images, $F(x^s)$. See Figure 3 for a diagram illustrating these interactions. The discriminators D_s and D_t have PatchGAN architecture [16], which only penalize local structures at the patch scale; therefore, they mainly capture local style statistics [22].

Our full objective includes three types of terms. First, we use two adversarial losses, one for matching the distributions of generated source style images, and of translated images from target to source, to the source data distribution. Similarly, the other is used for matching the distributions of generated target style images, and of translated images from source to target, to the target data distribution. Second, we use cycle-consistency losses to regularize the mapping functions F and H. Third, we use autoencoder losses between the generated paired style images.

4.1. Adversarial Losses

To learn the source discriminator D_s we use the typical adversarial loss [11] as follows:

$$\begin{aligned}\mathcal{L}_{adv-s}(D_s, G_s, H) &= E[\log(D_s(X^s))] \\ &+ E[\log(1 - D_s(H(X^t)))] + E[\log(1 - D_s(G_s(Z)))] \,,\end{aligned} \tag{1}$$

where G_s tries to generate images, from the random variable Z, that have local style statistics similar to source images X_s (and are paired with target style generated images $f_t = G_t(z)$). H tries to translate images from target into images with style similar to the source. D_s aims

Figure 5: **Black hair-to-blond hair and blond hair-to-black hair.** For every triple the left image is from the source, the middle image is reconstructed after translation, and the third one is the translated image.

at matching the distributions of generated source style images ($f_s = G_s(z)$), and of translated images from target to source ($H(x^t)$), to the source data distribution (x^s). H and G_s try to minimize (1) against the adversary D_s that tries to maximize it.

Similarly, to learn the target discriminator D_t we use the loss:

$$\begin{aligned}&\mathcal{L}_{adv-t}(D_t, G_t, F) = E[\log(D_t(X^t))] \\ &+ E[\log(1 - D_t(F(X^s))) + E[\log(1 - D_t(G_t(Z)))] \,,\end{aligned} \tag{2}$$

where G_t tries to generate images from Z that have local style statistics similar to target images X_t (and are paired with source style generated images $f_s = G_s(z)$). F tries to translate images from source into images with style similar to the target. D_t aims at matching the distributions of generated target style images ($f_t = G_t(z)$), and of translated images from source to target ($F(x^s)$), to the target data distribution (x^t). F and G_t try to minimize (2) against the adversary D_t that tries to maximize it.

Since D_s and D_t penalize only local structures, the loss associated with the discriminators D_s and D_t can be interpreted as a form of texture loss, which we loosely speaking refer to as "style".

4.2. Cycle Consistency Loss

By playing an adversarial game with (1) and (2) we could learn D_s, D_t, and also the translation networks F and H. However, doing so would not give sufficient guarantees that x_s and $\hat{x}_t$ were paired up in a meaningful ways. Indeed, the main requirement satisfied so far would be that the distribution of the translations matches the one of the target domain data. To alleviate that the translation networks should be cycle-consistent [43], meaning that the original image after translation should be reconstructable by a backward translation. This means that for source samples we want to have $x_s \xrightarrow{F} \hat{x}_t \xrightarrow{H} x_s$, while for target samples we want to have $x_t \xrightarrow{H} \hat{x}_s \xrightarrow{F} x_t$. Therefore, we adopt the following cycle-consistency loss:

$$\begin{aligned}\mathcal{L}_{cyc}(F, H) =& E_{X^s}[\|H(F(x_s)) - x_s\|_1] \\ &+ E_{X^t}[\|F(H(x_t)) - x_t\|_1] \,,\end{aligned} \tag{3}$$

Even after imposing (3), nothing would prevent the semantic content of images from changing during the translation. For example, a source image depicting a digit could still be translated into a target image depicting a different digit (e.g., where the source training dataset could be MNIST [21] and the target dataset could be SVHN [32]).

4.3. Autoencoder Loss

This loss allows to directly exploit the generation of paired style images for improving the learning of the translation networks. Specifically, the adversarial losses (1) and (2) allow us to generate source and target style images. In addition, those are paired by virtu of the fact that G_s and G_t share the initial layers [26], as described in detail in Section 4.5. Such paired style images can then be used to improve the training of F and H. More precisely, let us assume that G_s and G_t generate a pair of images (f_s, f_t) by drawing z. Since $f_t = G_t(z)$ is paired with $f_s = G_s(z)$, it should be possible to reconstruct f_t by translating f_s with F, i.e., $f_s \xrightarrow{F} f_t$. Similarly, since $f_s = G_s(z)$ is paired with $f_t = G_t(z)$, it should be possible to reconstruct f_s by translating f_t with H, i.e., $f_t \xrightarrow{H} f_s$. Therefore, we propose to add the following autoencoder loss to enforce what

we have just observed:

$$\mathcal{L}_{auto}(F,H) = E_Z[\|F(G_s(z)) - G_t(z)\|_1 + \|H(G_t(z)) - G_s(z)\|_1] \,, \quad (4)$$

The main effect of this loss should be to improve the way in which translation is performed by better learning how local statistics from the source domain should be mapped, or translated, onto the target domain. Since this training boost is focused only on the local statistics, i.e., the style of the domain, adding (4) should further discourage the mutation of structures at larger scales during the translation, which could potentially alter the semantic content of the images. This loss teaches the translation networks where to attend which makes it unique compared to [24].

4.4. Full Objective

The full objective is the summation of the losses described above:

$$\mathcal{L}(H,F,G_s,G_t,D_s,D_t) = \mathcal{L}_{adv-s}(D_s,G_s,H) + \mathcal{L}_{adv-t}(D_t,G_t,F) + \lambda_c\mathcal{L}_{cyc}(F,H) + \lambda_a\mathcal{L}_{auto}(F,H) \,, \quad (5)$$

where λ_c and λ_a are hyperparameters that strike a balance between the different losses. The desired translation networks can be found by solving:

$$F^*, H^* = \arg \min_{F,H,G_s,G_t} \max_{D_x,D_y} \mathcal{L}(F,H,G_s,G_t,D_s,D_t) \,. \quad (6)$$

4.5. Implementation Details

Each mapping network (F and H) includes 2 down-sampling convolutional layers followed by 9 residual blocks [13] and 2 up-sampling convolutional layers. We designed the number of kernels to have 256 feature maps in the residual blocks. Therefore, the noise dimension is $256 \times M \times M$, where M is one fourth of the input image dimension. G_s and G_t include 4 residual blocks and 2 up-sampling convolutional layers and share all the layers except the last convolutional layer to guarantee the generated pairs as discussed in [26]. We use PatchGan [17] and [44] because of its ability to capture local style statistics.

5. Experiments

We perform several qualitative and quantitative experiments. For the qualitative results, we perform several popular image-to-image translation tasks, including zebra-to-horse, black hair-to-blond hair, and apple-to-orange. For the quantitative experiments, we designed a number of domain adaptation tasks.

5.1. Qualitative results

Zebra-to-horse. We used around 1000 images of horses and 1300 images of zebras obtained from the ImageNet dataset [34]. All images are resized to 256×256. Figure 4a shows some horse images and their translation to zebras. We can even see in Figure 4a in the top-left and 2nd row left if we look at the water section of the images that the model, because of our focus on texture translation, has the capability of recognizing the reflections of the horse(s) and converting them also. We would not be able to learn this in supervised settings since we could never construct such an image pair. Further, this shows that due to learning texture images our model is generally spatially invariant, including rotation in image. Figure 4b and 6 show the translation from zebras to horse and some style pair images respectively.

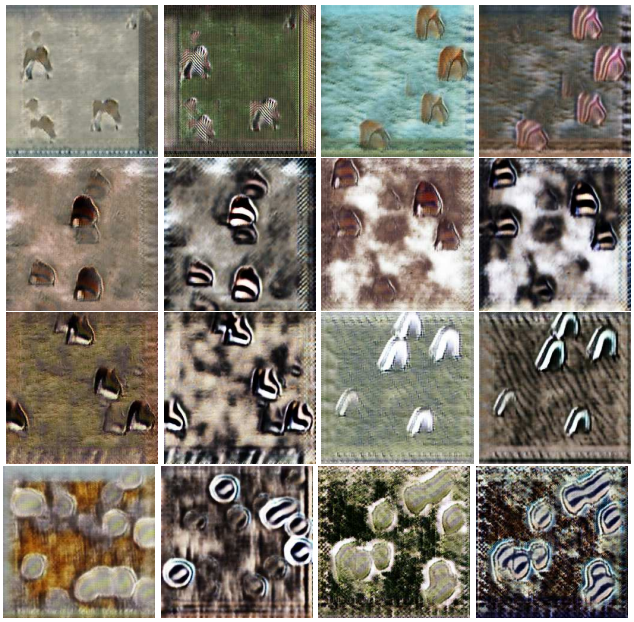


Figure 6: **Paired style images from/to Zebra to/from Horse.** We train the translation networks F and H using the generated paired style images by adding an autoencoder loss. Each zebra style image contains several striped texture and its paired horse style image contains more uniform texture in the same exact positions. By training the mapping networks using these paired style images, we teach them to look for and translate only the desired style that emerges by contrasting the source with the target domains.

Black hair-to-blond hair. We used 2000 images of women with black hair and women with blond hair obtained from the CelebA dataset [27]. All images are resized to 128×128. Figure 5 shows some translated images from black hair to blond hair and vice versa. Because of the reduced image size, the PatchGAN discriminator can also capture the global style statistics in addition to local style statistics, leading to generated images that look more realistic. See Figure 8.

Apple-to-orange. We used around 1000 images of apples and oranges obtained from ImageNet dataset [34]. All images are resized to 256×256. Figure 7 shows some samples and Figure 2b shows some generated styles.

Figure 7: **Apple-to-orange and orange-to-apple.** For every triple the left image is from the source, the middle image is reconstructed after translation, and the third one is the translated image.

Figure 8: **Paired style images from/to black hair to/from blond hair.** Since we resize the images to 128×128, the PatchGAN discriminator can also capture the global style statistics in addition to local style statistics. That is the reason why some generated images look realistic. By training the mapping networks using these paired images, we teach them to look only for and translate only the desired style that emerges by contrasting the source with the target domains.

5.2. Quantitative results - Domain Adaptation

Preserving the semantics in images during translation is one of the main advantages of our proposed model. In this section, we design an experiment that highlights this aspect.

Domain Adaptation. Deep learning approaches have shown promising results when large amounts of labeled data are available. There are still many problems worth solving where labeled data on an equally large scale is too expensive to collect, annotate, or both. In such a scenario, the typical approach is to train a model from a closely related labeled dataset (*source* dataset with N labeled samples (x_s^i, y_s^i), $i = 1, \cdots N$) with a large amount of samples and adapt it to work well on the *target* domain, the one that has no or few labeled samples. Domain adaptation usually is done in three ways. Some try to find a network that maps from the source domain to the target domain [24]. Some find a shared latent space that both domains can be mapped to before classification [31, 39, 30]. Finally, several works use regularization to improve the fit on the target domain [2, 1]. The first group usually outperforms the others if mapping networks are able to preserve the semantics. We follow the same strategy and our goal is to translate the source images, which are labeled, to the target domain using our mapping network F. Assuming that F does not change the semantic content, the labels are preserved, and it is possible to train a classifier on the translated images $(\hat{x}_t^i = F(x_s^i), y^i)$.

Table 1: **Domain adaptation.** Classification accuracy for domain adaptation over the 100 classes of $\mathcal{M}, \mathcal{U}$, and $\mathcal{S}$.

Method	*LB*	*CoGAN* [26]	*UNIT* [24]	*CycleGAN* [43]	*Ours*	*UB*
$\mathcal{M} \rightarrow \mathcal{U}$	82.2	95.6	95.9	95.6	95.8	96.3
$\mathcal{U} \rightarrow \mathcal{M}$	69.6	93.1	93.5	96.4	94.5	99.2
$\mathcal{S} \rightarrow \mathcal{M}$	67.1	-	90.5	70.3	89.9	99.2

The MNIST ($\mathcal{M}$), USPS ($\mathcal{U}$), and SVHN ($\mathcal{S}$) datasets have been widely used for domain adaptation [8, 33, 39]. We considered three cross-domain tasks. They include $\mathcal{M} \rightarrow \mathcal{U}$, $\mathcal{U} \rightarrow \mathcal{M}$, $\mathcal{S} \rightarrow \mathcal{M}$. We followed the experimental setting in [24], which involves using the full training sets during learning phases and evaluation on the standard test sets. Table 1 shows the results of our model compared with the state-of-the-art. In Table 1, *LB* stands for lower bound when we train the LeNet architecture with source samples and test on target samples (no adaptation). *UB* stands for upper bound when we train the LeNet architecture with target samples and test on target samples. As the table shows, the performance of the state-of-the-art methods are very close to the upper bound and our model shows to compare well. In particular, it performs significantly better than CycleGAN on the task $\mathcal{S} \rightarrow \mathcal{M}$, which seems to suffer more than in the others the problem of flipping the semantic content during the translation.

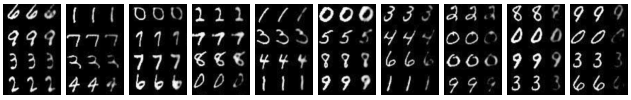


Figure 9: **MNIST.** Each triplet contains an MNIST sample x_s, its reconstructed image $H(F(x_s))$, and its translation to the USPS domain $F(x_s)$. The proposed autoencoder loss prevents label flipping during the translation.

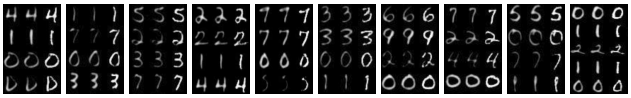


Figure 10: **USPS.** Each triplet contains an USPS sample x_t, its reconstructed image $F(H(x_t))$, and its translation to the MNIST distribution $H(x_t)$. The proposed autoencoder loss prevents label flipping during the translation.

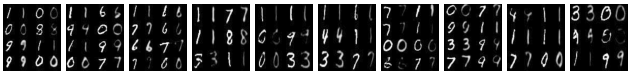
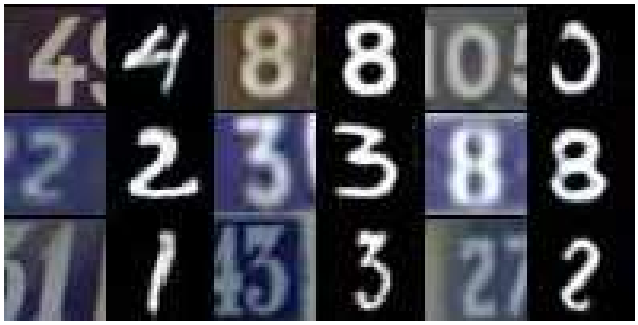


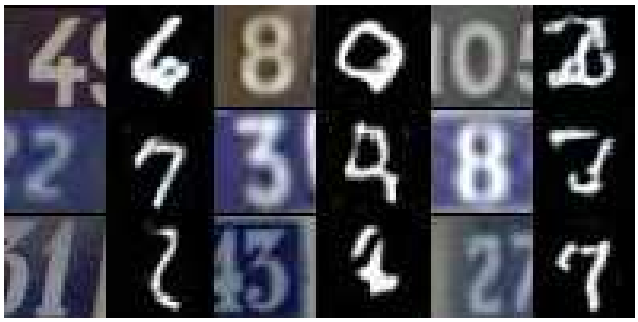
Figure 11: **Paired digits generation.** Each pair contains a generated image from MNIST and a generated image from the USPS distributions.

Figure 9 shows some translation examples from MNIST to USPS. In particular, it shows several triplets containing an original image x_s, its reconstructed image $H(F(x_s))$, and its translation $F(x_s)$. The translated images look very similar to the USPS samples. Figure 10 shows several triplets containing an original image from USPS x_t, its reconstructed image $F(H(x_t))$, and its translation $H(x_t)$. The translated images look very similar to the MNIST samples. Figures 9 and 10 show that there is no *label flipping* during the translation. Figure 11 shows some generated samples using G_s and G_t. Since the image dimension is small in the case of digits (28×28), we use a simple discriminator with one output neuron. Therefore, the generated images look similar to the training samples (not their style).

Semantic flipping. It is instructive to see what goes wrong during the translation process from SVHN to MNIST (see $\mathcal{S} \rightarrow \mathcal{M}$ in Table 1) of CycleGAN and compare it with the proposed approach. That comparison is depicted in Figure 12, where samples from the SVHN dataset are translated into the MNIST domain using our approach in Figure 12a, and the same samples are translated using CycleGAN in Figure 12b. For these SVHN samples we note that CycleGAN is having some difficulty in translating the images without causing enough distortion that will then lead to a classification error, or even without generating an image digit that has a semantic meaning that is visibly different from the semantic meaning of the input sample. For instance the input number 4 is translated into a digit that looks like a 6 or the number 2 is translated into an image that looks like a 7. This level of distortion or even semantic flipping is not present for the translation of the same SVHN samples, as shown in Figure 12a. This difference in the ability to do the translation is responsible for the significant drop in accuracy of CycleGAN, while our approach is able to compare well and often exceed other approaches.



(a) Proposed approach



(b) CycleGAN

Figure 12: **Translation comparison - SVHN2MNIST.** Pairs of images depicting, on the left, samples from the SVHN dataset, and on the right, the same samples translated into the target domain distribution defined by MNIST. The top set of images, (a), has been produced with the proposed approach. The bottom set of images, (b), has been produced with CycleGAN.

6. Conclusion

We have proposed an unsupervised image-to-image translation model that leverages paired style image generation to improve the preservation of semantic content, which we have shown is both effective for translation and in specific tasks such as domain adaptation. The semantic preservation experiments and the image translation results show that our method compares well with the state-of-the-art, while being reliant on simple, more computationally and memory efficient structures than those which rely on style multi-stage generation [42, 19]. In the future we intend to explore what changes to the model would be necessary for it to yield useful results on translation tasks such as super-resolution or depth-estimation. We also intend to explore the inspiration from style transfer of leveraging the Gram matrix to see if it will further improve the learning of the local style statistics.

References

[1] Carlos J Becker, Christos M Christoudias, and Pascal Fua. Non-linear domain adaptation with boosting. In *Advances in Neural Information Processing Systems*, pages 485–493, 2013.

[2] Alessandro Bergamo and Lorenzo Torresani. Exploiting weakly-labeled web images to improve object classification: a domain adaptation approach. In *Advances in Neural Information Processing Systems*, pages 181–189, 2010.

[3] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised Pixel-Level Domain Adaptation with Generative Adversarial Networks. *CVPR*, pages 3722–3731, 2016.

[4] X Chen, C. Xu, X. Yang, and D. Tao. Attention-gan for object transfiguration in wild images. In *European Conference on Computer Vision (ECCV)*, 2018.

[5] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. *arXiv preprint arXiv:1711.09020*, 2017.

[6] Alexei A. Efros and William T. Freeman. Image quilting for texture synthesis and transfer. *Proceedings of the 28th annual conference on Computer graphics and interactive techniques - SIGGRAPH '01*, (August):341–346, 2001.

[7] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *IEEE ICCV*, pages 2960–2967, 2013.

[8] B. Fernando, T. Tommasi, and T. Tuytelaarsc. Joint cross-domain classification and subspace learning for unsupervised adaptation. *Pattern Recogition Letters*, 2015.

[9] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. *The IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016.

[10] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2066–2073. IEEE, 2012.

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014.

[12] R. Gopalan, R. Li, and R. Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *IEEE ICCV*, pages 999–1006, 2011.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] Aaron Hertzmann, Charles E. Jacobs, Nuria Oliver, Brian Curless, and David H. Salesin. Image analogies. *Proceedings of the 28th annual conference on Computer graphics and interactive techniques - SIGGRAPH '01*, (August):327–340, 2001.

[15] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A. Efros, and Trevor Darrell. CyCADA: Cycle-Consistent Adversarial Domain Adaptation. pages 1–15, 2017.

[16] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *arXiv preprint arXiv:1611.07004*, 2016.

[17] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-Image Translation with Conditional Adversarial Networks. *CVPR*, pages 1125–1134, 2017.

[18] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. *Eccv*, pages 694–711, 2016.

[19] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[20] Brian Kulis, Kate Saenko, and Trevor Darrell. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1785–1792. IEEE, 2011.

[21] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[22] Chuan Li and Michael Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *European Conference on Computer Vision*, pages 702–716. Springer, 2016.

[23] Alexander H Liu, Yen-Cheng Liu, Yu-Ying Yeh, and Yu-Chiang Frank Wang. A unified feature disentangler for multi-domain image translation and manipulation. In *Advances in Neural Information Processing Systems*, pages 2590–2599, 2018.

[24] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems*, pages 700–708, 2017.

[25] Ming-Yu Liu and Oncel Tuzel. Coupled Generative Adversarial Networks. *Advances in Neural Information Processing Systems {(NIPS)}*, (Nips):469–477, 2016.

[26] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. In *Advances in Neural Information Processing Systems*, pages 469–477, 2016.

[27] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2015.

[28] Fujun Luan, Sylvain Paris, Eli Shechtman, and Kavita Bala. Deep Photo Style Transfer. 2017.

[29] Youssef Alami Mejjati, Christian Richardt, James Tompkin, Darren Cosker, and Kwang In Kim. Unsupervised attention-guided image-to-image translation. In *Advances in Neural Information Processing Systems*, pages 3693–3703, 2018.

[30] Saeid Motiian, Quinn Jones, Seyed Iranmanesh, and Gianfranco Doretto. Few-shot adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, pages 6673–6683, 2017.

[31] Saeid Motiian, Marco Piccirilli, Donald A. Adjeroh, and Gianfranco Doretto. Unified deep supervised domain adaptation and generalization. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[32] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, 2011.

[33] Artem Rozantsev, Mathieu Salzmann, and Pascal Fua. Beyond sharing weights for deep domain adaptation. *arXiv preprint arXiv:1603.06432*, 2016.

[34] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 2015.

[35] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *ECCV*, pages 213–226, 2010.

[36] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial training. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[37] Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised Cross-Domain Image Generation. *International Conference on Learning Representations (ICLR)*, pages 1–14, 2017.

[38] Tatiana Tommasi, Martina Lanzi, Paolo Russo, and Barbara Caputo. Learning the roots of visual domain shift. In *Computer Vision–ECCV 2016 Workshops*, pages 475–482. Springer, 2016.

[39] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[40] Daan Wynen, Cordelia Schmid, and Julien Mairal. Unsupervised learning of artistic styles with archetypal style analysis. In *Advances in Neural Information Processing Systems*, pages 6584–6593, 2018.

[41] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. DualGAN: Unsupervised Dual Learning for Image-to-Image Translation. *Proceedings of the IEEE International Conference on Computer Vision*, 2017-October:2868–2876, 2017.

[42] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaolei Huang, Xiaogang Wang, and Dimitris Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. *arXiv preprint arXiv:1612.03242*, 2016.

[43] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[44] Jun Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. *Proceedings of the IEEE International Conference on Computer Vision*, 2017-Octob:2242–2251, 2017.