Learning Unsupervised Cross-domain Image-to-Image Translation using a Shared Discriminator

Rajiv Kumar[a], Rishabh Dabral and G. Sivakumar[b]
CSE Department, Indian Institute of Technology Bombay, Mumbai, India

Keywords: Image-to-Image Translation, Unsupervised Learning, Cross-domain Image Translation, Shared Discriminator, Generative Adversarial Networks.

Abstract: Unsupervised image-to-image translation is used to transform images from a source domain to generate images in a target domain without using source-target image pairs. Promising results have been obtained for this problem in an adversarial setting using two independent GANs and attention mechanisms. We propose a new method that uses a single shared discriminator between the two GANs, which improves the overall efficacy. We assess the qualitative and quantitative results on image transfiguration, a cross-domain translation task, in a setting where the target domain shares similar semantics to the source domain. Our results indicate that even without adding attention mechanisms, our method performs at par with attention-based methods and generates images of comparable quality.

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS

1 INTRODUCTION

Generative Adversarial Networks(GANs) (Goodfellow et al., 2014) belong to the class of generative models (Kingma and Welling, 2013) widely used in various image generation and translation tasks like computer vision and image processing (Johnson et al., 2016), (Wang et al., 2019), (Wu et al., 2017). While the state-of-the-art methods (Huang et al., 2018), (Liu et al., 2019), (Park et al., 2019) in image-to-image translation tasks are significantly good (Wang et al., 2018b), (Mejjati et al., 2018), (Tang et al., 2019) across multi-domain and cross-domain tasks, there is still room for improvement in image transfiguration tasks. Most of the image-to-image translation tasks assume the availability of source-target image pairs (Isola et al., 2017), (Zhu et al., 2017b) or expect the source-target pairs to have rough alignment between them (Isola et al., 2017), (Wang et al., 2018b). However, there are scenarios where source-target image pairs are not available or when arbitrarily selected source-target image pairs have poor alignment between them.

While most image-to-image translation tasks involve translation over the complete image, there are cases where only an object of interest needs to be translated in the source and target domain. Let's consider the case of translating images of apples to oranges or horses to zebras. In both cases, only the object of interest needs to be translated, without affecting the rest of the image or it's background. This calls for the need of attention mechanisms (Kastaniotis et al., 2018), (Qian et al., 2017), (Zhang et al., 2018), (Talreja et al., 2019) to attend to the objects of interest. Contrast-GAN (Liang et al., 2017) is a work that has used object-mask annotations to guide the translation at high-level semantic levels at the cost of extra data. However, recent works have used attention mechanisms (Wang et al., 2017), (Mejjati et al., 2018), (Tang et al., 2019) without using any extra data or pretrained models. Moreover, very few works focus on image transfiguration, a cross-domain image-to-image translation task in an unsupervised setting without using additional networks, extra data or attention mechanisms.

In this paper, we focus on the above problem by proposing a framework that unifies the capabilities of multiple discriminators into a shared one, which not only improves the efficacy but also works without using extra data(object masks) or attention mechanisms. Adversarial training of the network involves combining the labels of the domains from different tasks conditioned on the input image and optimizing the objectives of the networks. We believe that there has

[a] https://orcid.org/0000-0003-4174-8587
[b] https://orcid.org/0000-0003-2890-6421

Kumar, R., Dabral, R. and Sivakumar, G.
Learning Unsupervised Cross-domain Image-to-Image Translation using a Shared Discriminator.
DOI: 10.5220/0010184102560264
In *Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2021) - Volume 5: VISAPP*, pages 256-264
ISBN: 978-989-758-488-6
Copyright © 2021 by SCITEPRESS – Science and Technology Publications, Lda. All rights reserved

not been any previous work where a dual generator shared discriminator setup has been used for **cross-domain image-to-image translation** and we are the first to propose a novel method. We summarize the paper contribution as follows:

1. We improve the efficacy of the GANs used for unsupervised cross-domain image-to-image translation tasks by introducing a novel shared discriminator setup. We empirically demonstrate the effectiveness of our method on image transfiguration tasks and report the qualitative and quantitative results on two datasets.
2. We conduct an ablation study to study the efficacy of the networks, training objectives and architectures keeping the dataset and other parameters constant and report the quantitative results of the study.

2 RELATED WORK

Generative Adversarial Networks. GANs are generative networks that use a trainable loss function to adapt to the differences between the data distributions of generated images and the real images. Since their inception (Goodfellow et al., 2014) (Radford et al., 2015), GANs have been used in various applications from computer vision (Ma et al., 2017), (Vondrick et al., 2016), image-to-image translation (Taigman et al., 2016), (Tung et al., 2017), video-to-video translation (Wang et al., 2019), (Wang et al., 2018a), image super-resolution (Ledig et al., 2016), etc. among others. We refer interested readers to read more about GANs from (Creswell et al., 2018),(Jabbar et al., 2020), (Kurach et al., 2018) and (Wang et al., 2020).

Image-to-Image Translation. Recent image-to-image translation works like pix2pix (Isola et al., 2017), pix2pixHD (Wang et al., 2018b) use conditional GANs to learn a mapping from source domain images to target domain images. While some rely on paired source-target images, works like CycleGAN, DualGAN, DiscoGAN (Kim et al., 2017) and (Tung et al., 2017), (Taigman et al., 2016), (Liu and Tuzel, 2016), (Liu et al., 2017), (Bousmalis et al., 2016) learn the mapping between the source domain and target domain without using any paired images. CoGAN (Liu and Tuzel, 2016) also learns the joint distribution of multi-domain images by sharing weights of generators and discriminators. UNIT (Liu et al., 2017) uses a shared latent space framework built on CoGANs to learn a joint distribution of different domain images and achieves very high quality image translation results.

In an adversarial setting, image-to-image translation involves generators that learn mappings to translate images from a source domain to a target domain and vice-versa. Furthermore, adversarial methods that involve GAN either share network weights (Liu and Tuzel, 2016), (Talreja et al., 2019) or use mechanisms (Yi et al., 2017), (Zhu et al., 2017a) that involve a primal GAN and a dual GAN. A Dual-GAN (or DualGAN) (Yi et al., 2017) setup employs two discriminators: a primal GAN and a dual GAN, performing inverse tasks of each other. Each discriminator is trained to discriminate target domain images as positive samples and translated source domain images as negative samples. Similarly, in CycleGAN (Zhu et al., 2017a), the primal-dual relation is regularized by a forward consistency loss and backward cycle consistency loss, which constitutes the cycle-consistency loss. This reduces the space of possible mappings by enforcing a strong relation across domains.

Conventionally, separate task-specific generators and discriminators are needed for image-to-image translation, since each network deals with a different set of real and fake images. However, StarGAN (Choi et al., 2018) achieves multi-domain image translation using a single generator by considering each domain as a set of images with a common attribute (for e.g. hair color, gender, age, etc.) and by exploiting the commonalities in the datasets. Similarly, a Dual Generator GAN(G^2GAN) (Tang et al., 2019) consists of two task-specific generators and single discriminator focusing on multi-domain image-to-image translation. However, their optimization objective is complex, consisting of five components including color consistency loss, MS-SSIM loss and conditional identity preserving loss for preventing mode collapse. While Dual Generator GAN uses a single discriminator, the underlying task is multi-domain image translation. However, in this paper we focus on the task of cross-domain image translation using a single shared discriminator.

3 METHODOLOGY

We briefly explain the problem formulation in subsection 3.1, proposed framework in subsection 3.2, image pools in subsection 3.2.1, training stages in subsection 3.3 and loss functions in subsection 3.4 below.

3.1 Problem Formulation

For the image-to-image translation problem, our goal is to learn two mapping functions, $G_{AB} : A \rightarrow B$ and

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

$G_{BA}: B \rightarrow A$, between domains A and B modelled by generators G_{AB} and G_{BA} respectively. We consider task-specific generators since the input distribution is different for each task in cross-domain image-to-image translation. A domain is referred to as either *source* or *target* domain, based on its role in the translation task. The goal of the generator G_{AB} is to translate an input image a from source domain A to the target domain, such that the generated image b^* follows the distribution of the target domain B, $p_{data}(B)$. Likewise, the task of generator G_{BA} is to translate an image $b \in B$ to an image a^* such that it follows the distribution of the target domain A, $p_{data}(A)$. We propose to provide adversarial supervision using a novel shared discriminator, D_{shared} common to both the generators without using extra networks, masks or additional data. In this paper, we focus our method on transfiguration tasks, which requires translation of objects of interest while keeping other objects and the background same. Some transfiguration tasks include apples ↔ oranges, horses ↔ zebras, etc.

3.2 Proposed Framework

Each translation task ($A \rightarrow B$ and $B \rightarrow A$) is mapped to a separate generator. For guided image generation, we use conditional GANs (Mirza and Osindero, 2014) that condition using the input images. During training, each generator learns to translate its input from *source* domain to the corresponding *target* domain. However, our approach differs from the conventional setting, which treats the target domain samples as *real* and translated images as *fake*. Instead, we exploit the fact that the data distributions of the source and the target domains of one translation task are the same as that of the target and the source domains of its inverse translation task.

In our novel formulation, the proposed shared discriminator D_{shared} is trained to classify the generated images into either belonging to domain A or domain B. The translated images and random images from the two domains are conditioned on the input images to form the base for adversarial training using the shared discriminator. We hypothesize that this unification allows for *domain-aware* discrimination which is crucial for tasks like transfiguration, where a specific part of the image with distinct feature sets are to be transformed. GANs are infamous for unstable training and prone to model oscillation. To stabilize the model training, we leverage the power of image pools with modifications tailored for our approach. Once the training is complete, the generator outputs are treated as final prediction and the discriminators are not needed in inference stage.

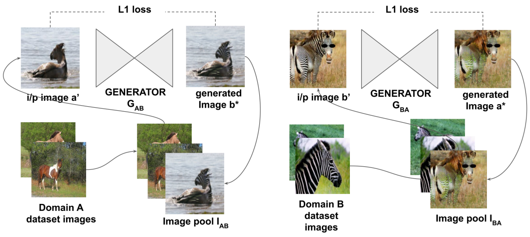

Figure 1: The above image corresponds to training stage 1, with two generators and two image pools. Here the generated images are pushed to the same image pool as that of the translation task. Shared discriminator has been avoided for brevity.

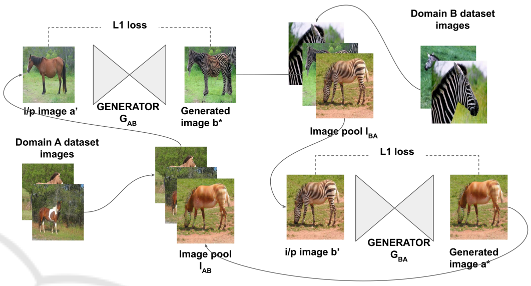

Figure 2: The above image corresponds to training stage 2, with two generators and two image pools. Here the generated images are pushed to the image pool of the corresponding inverse translation task. Shared discriminator has been avoided for brevity.

3.2.1 Image Pools

Generally, the generator outputs are reused in image-to-image translation techniques that involves a reconstruction loss between the source image and the reconstructed image(the resulting image after undergoing two translations, from source domain to the target domain and back to the source domain). An image pool (Shrivastava et al., 2016) is generally used to store a history of generated images to feed the discriminator in order to reduce the model oscillation during adversarial training. In our method, we associate an image pool to the generator of each translation task, such that the translated images can be reused as inputs to either of the generators by pushing to one image pool or the other, i.e. image pool I_{AB} is associated with G_{AB} and image pool I_{BA} is associated with G_{BA} (see Fig. 1 and Fig. 2). We use this simple tweak to improve the robustness of the generators to deal with variety of input images. In some cases, we also observe performance improvements, which we discuss later in the ablation study. Since the generated images are pushed to the image pools, each image pool gets a static input set of source domain images and an evolving input set of generated images

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

from one generator or the other, depending upon the training stage.

3.3 Training Stages

We consider the image-to-image translation task $A \leftrightarrow B$ by learning the translation mapping $G_{AB} : A \rightarrow B$ and it's inverse translation mapping $G_{BA} : B \rightarrow A$. Throughout the training process, the inputs to the generators G_{AB} and G_{BA} are from the image pools $I_{AB} = \{a'_1, a'_2, \dots, a'_{|I_{AB}|}\}$ and $I_{BA} = \{b'_1, b'_2, \dots, b'_{|I_{BA}|}\}$ respectively. The image pools are initialized by the images of their source domains, A and B. The details of each training stage are given below.

Training Stage 1. If we consider the initial stages of training, the translated images appear closer in appearance to the source domain with very few target domain features. Therefore, we interleave the translated images from the generator with the source domain images using the same image pool, i.e. I_{AB} would pool-in images a from A and $b*$ from $G_{AB}(a)$, and I_{BA} would pool-in images b from B and $a*$ from $G_{BA}(b)$ as depicted in Fig. 1.

Training Stage 2. As training proceeds, the generators improve upon their translation capabilities and the generated images possess more target domain features and very few source domain features. Therefore, each generator can take the outputs of the other generator as adversarial images, in addition to their respective source domain images, i.e. I_{AB} would pool-in images a from A and $a*$ from $G_{BA}(b)$, and I_{BA} would pool-in images b from B and $b*$ from $G_{AB}(a)$ as depicted in Fig. 2. The generated images are pushed to the image pool of the inverse translation task, to mimic cyclic translations as done in some related works.

3.4 Loss Functions

Conventionally, a discriminator is used to distinguish between *real* images from the dataset and *fake* images generated by the generator. However, we avoid the usage of the terms *real* images and *fake* images, and use abstract binary labels *True* and *False* instead. We assign the same labels for a domain irrespective of the translation task or their role in the translation task, i.e. the label assigned for the source domain images in the forward translation is the same as that of the target domain images in the inverse translation task. We assign *true* labels for domain B images and *false* labels for domain A images.

Discriminator Loss. All translated images are conditioned on the their input images when subjected to the discriminator, while optimizing the objectives of D_{shared}, i.e. b^* is conditioned on a' and a^* is conditioned on b'. The generated images, a^* from $G_{AB}(a')$ and b^* from $G_{BA}(b')$ are labelled the same labels as their source domain images a and b respectively, while subjecting to the discriminator. The shared discriminator D_{shared} is trained with a binary cross entropy loss $L_{D_{shared}}$. The goal of D_{shared} is to classify the generated images into either domain A or domain B depending upon the source domain of the translation task. In addition, we subject the shared discriminator to random domain B images labelled as *true* and random domain A images labelled as *false*. These random images are conditioned on input images a' or b' depending on the translation task or whether they are input to generator G_{AB} or G_{BA} respectively. Formally, the complete training objective of D_{shared} or the discriminator loss function is given by,

$$\begin{aligned}
L_{D_{shared}}(G_{AB}, G_{BA}, D_{shared}, A, B, I_{AB}, I_{BA}) = \\
E_{b\sim p_{\text{data}}(b)}[log(D_{shared}(b|a'))] + \\
E_{a\sim p_{\text{data}}(a)}[log(1 - D_{shared}(a|a'))] + \\
E_{a'\sim p_{\text{data}}(a')}[log(1 - D_{shared}(G_{AB}(a')|a'))] + \\
E_{a\sim p_{\text{data}}(a)}[log(1 - D_{shared}(a|b'))] + \\
E_{b\sim p_{\text{data}}(b)}[log(D_{shared}(b|b'))] + \\
E_{b'\sim p_{\text{data}}(b')}[log(D_{shared}(G_{BA}(b')|b'))]. \qquad (1)
\end{aligned}$$

The first three parts of Eq. 1 are conditioned on input images a' from image pool I_{AB} and represent the translation $A \rightarrow B$, while the latter parts are conditioned on the input images b' from image pool I_{BA} and represent the translation $B \rightarrow A$.

Generator Loss. We enforce a reconstruction loss between the generator's input and it's output involving only one image translation, in contrast to conventional pixel reconstruction objectives that involves translations over both directions. We choose a loss function that can preserve the median values, so that the objects of interest are translated without translating other objects in the image or the background. This motivates the use of L_1 pixel reconstruction loss between the input and output of each generator with additional help from adversarial training. The adversarial goal of each generator is to fool the shared discriminator into identifying generated images as belonging to the target domain images, i.e. G_{AB} tries to map b^* as belonging to B while G_{BA} tries to map a^* as belonging to A. The adversarial losses overrule the reconstruction loss over the membership score of the generated image, which results in the source images to take target domain features. We can express the full objective of G_{AB} as the sum of Eq. 2, which corresponds to the adversarial loss and Eq. 3, which corresponds to the L_1 reconstruction loss. Similarly,

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS

we can express the full objective of G_{BA} as the sum of Eq. 4, which corresponds to the adversarial loss and Eq. 5, which corresponds to the L_1 reconstruction loss.

$$L_{G_{\text{AB}}}(G_{AB}, D_{shared}, I_{AB}) = E_{a' \sim p_{\text{data}}(a')}[log(D_{shared}(G_{AB}(a')))]. \tag{2}$$

$$L^{G_{\text{AB}}}_{pixel}(G_{AB}, I_{AB}) = E_{a' \sim p_{\text{data}}(a')}[\|G_{AB}(a') - a'\|_1]. \tag{3}$$

$$L_{G_{\text{BA}}}(G_{BA}, D_{shared}, I_{BA}) = E_{b' \sim p_{\text{data}}(b')}[log(1 - D_{shared}(G_{BA}(b')))]. \tag{4}$$

$$L^{G_{\text{BA}}}_{pixel}(G_{BA}, I_{BA}) = E_{b' \sim p_{\text{data}}(b')}[\|G_{BA}(b') - b'\|_1]. \tag{5}$$

4 IMPLEMENTATION

We trained the tasks on 128x128 size images as well as on 256x256 size images. For training, the training images were resized to 1.125 times and were randomly cropped to the required size. The batch size for all our experiments was 4. Smaller batch sizes enable training with larger image sizes. Also, the image pools could be stored in the main memory or cuda device memory. We experimented with the *Adam* optimizer as well as *RMSProp*, and found that *Adam* gives better performance for most of our experiments. We used a learning rate of 0.0001 with the *Adam* optimizer with betas of 0.5 and 0.999. We used the adversarial loss for membership score with the vanilla GAN or binary cross entropy with logit loss. We used a lambda of 10.0 for the adversarial losses and a lambda in [100.0, 200.0] for the reconstruction loss.

4.1 Architecture

We use identical network architecture for both the generators throughout an experiment. We conduct experiments with the Resnet (He et al., 2015) architecture as well as Unet (Ronneberger et al., 2015) architecture. While using the Unet architecture, the generator has the same number of downsampling layers and upsampling layers with a bottleneck in between and skip connections connecting the downsampling and upsampling layers. Our proposed method doesn't use noise vectors as in the pix2pix implementation(Isola et al., 2017). Also, using dropout doesn't affect the performance of our method when implemented with the Unet architecture. In the Resnet architecture, the skip connections exist between Resnet blocks. The discriminator's architecture used in our experiments is PatchGAN (Zhu et al., 2017a).

Table 1: Effect of network architectures (Ronneberger et al., 2015) and (He et al., 2015) on translation tasks horse ↔ zebra and apples ↔ oranges. The results are compared using FID and KID scores.

FID	**Horse**	**Zebra**	**Apples**	**Oranges**
Unet	211.76 ± 3.65	119.99 ± 14.01	**164.87 ± 4.20**	172.30 ± 2.33
Resnet	**210.37 ± 5.10**	**97.47 ± 7.85**	168.86 ± 3.20	172.30 ± 2.33
KID	**Horse**	**Zebra**	**Apples**	**Oranges**
Unet	0.063± 0.002	0.046±0.003	**0.051 ± 0.003**	0.044± 0.002
Resnet	**0.058 ± 0.002**	**0.030 ± 0.002**	0.052 ± 0.002	0.044 ± 0.002

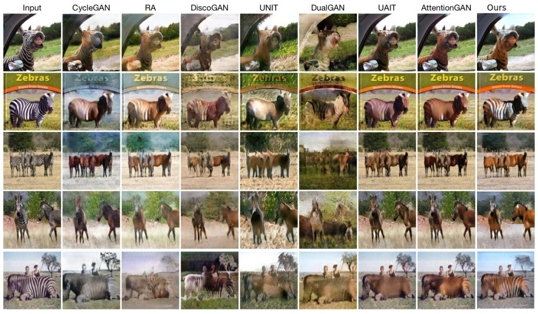

Figure 3: Comparison of test images generated by different methods (Zhu et al., 2017a), (Wang et al., 2017), (Kim et al., 2017), (Liu et al., 2017), (Yi et al., 2017), (Mejjati et al., 2018), (Tang et al., 2019) (left to right) on zebra to horse task. Leftmost column shows the input, rightmost column shows results from our method.

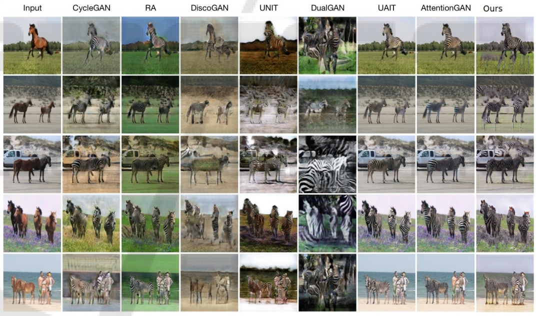

Figure 4: Comparison of test images generated by different methods (Zhu et al., 2017a), (Wang et al., 2017), (Kim et al., 2017), (Liu et al., 2017), (Yi et al., 2017), (Mejjati et al., 2018), (Tang et al., 2019) (left to right) on horse to zebra task. Leftmost column shows the input, rightmost column shows results from our method.

5 EXPERIMENTS AND EVALUATION

5.1 Datasets

We used *apples* to *oranges* dataset and *horse* to *zebra* dataset which were originally used in CycleGAN (Zhu et al., 2017a). These images are available from Imagenet with a training set size of each class having 939 (horse), 1177 (zebra), 996 (apple), and 1020 (orange) images.

Table 2: FID scores between generated samples and target samples for horse to zebra translation task on methods (Liu et al., 2017), (Zhu et al., 2017a), (Yang et al., 2018), (Tang et al., 2019) (from top to bottom). For this metric, lower is better.

Method	Horse → Zebra
UNIT	241.13
CycleGAN	109.36
SAT (Before Attention)	98.90
SAT (After Attention)	128.32
AttentionGAN	**68.55**
Ours	92.91

5.2 Evaluation Metric

We use the Frechet Inception Distance(FID) (Heusel et al., 2017) and Kernel Inception Distance(KID) (Bińkowski et al., 2018) preferably over metrics like Inception score. For both metrics, lower scores imply similarities in features between the compared sets of images. However, both metrics are adversely affected by the presence of adversarial noise and hallucinated features in the generated images that these metrics do not correlate to the judgement by human perception. This suggests that either metrics aren't better than each other, and better scores doesn't always imply better translation results. Hence, we consider those FID and KID scores from our experiments which are positively correlated.

Table 3: KID × 100 ± std. × 100 compared for different methods (Kim et al., 2017), (Wang et al., 2017), (Yi et al., 2017), (Liu et al., 2017), (Zhu et al., 2017a), (Mejjati et al., 2018), (Tang et al., 2019)(from left to right). Abbreviations: (H)orse, (Z)ebra (A)pple, (O)range.

Method	H → Z	Z → H	A → O	O → A
DiscoGAN	13.68 ± 0.28	16.60 ± 0.50	18.34 ± 0.75	21.56 ± 0.80
RA	10.16 ± 0.12	10.97 ± 0.26	12.75 ± 0.49	13.84 ± 0.78
DualGAN	10.38 ± 0.31	12.86 ± 0.50	13.04 ± 0.72	12.42 ± 0.88
UNIT	11.22 ± 0.24	13.63 ± 0.34	11.68 ± 0.43	11.76 ± 0.51
CycleGAN	10.25 ± 0.25	11.44 ± 0.38	8.48 ± 0.53	9.82 ± 0.51
UAIT	6.93 ± 0.27	8.87 ± 0.26	6.44 ± 0.69	5.32 ± 0.48
AttentionGAN	**2.03** ± **0.64**	6.48 ± 0.51	10.03 ± 0.66	**4.38** ± **0.42**
Ours	3.00 ± 0.20	**5.80** ±**0.20**	**4.40** ± **0.20**	5.10 ± 0.30

5.3 Experiments

We compute the KID score over 100 iterations and return its mean, while the FID scores are computed over 10 iterations and the mean value is returned. We compute the KID scores and FID scores on the test data using the generator models from the same checkpoint. We trained the tasks on 128x128 size images as well as on 256x256 size images and tested both category of models on 256x256 test images. We refer (Tang et al., 2019) to compile the experimental results in the qualitative comparisons and metric scores in Table 3 and 2. We report the performance comparison of different architectures on translation tasks *horse* ↔ *zebra* and *apples* ↔ *oranges*, measured in FID and KID scores in Table 1. We report the FID scores on horse → zebra translation task in Table 2. KID scores are compared over the horses ↔ zebras task and apples ↔ oranges task in Table 3. The results of qualitative comparisons includes the comparison of translated images from *zebras* → *horses* task in Fig. 3, *horses* → *zebras* task in Fig. 4, *oranges* → *apples* task in Fig. 5, and *apples* → *oranges* task in Fig. 6.

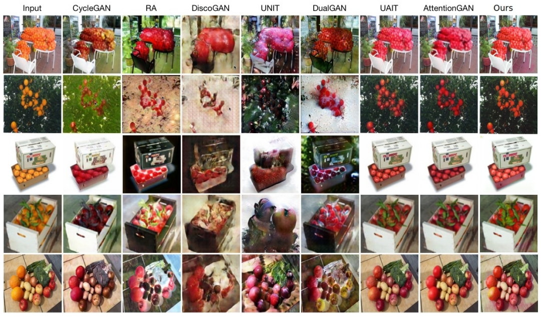

Figure 5: Comparison of test images generated by different methods (Zhu et al., 2017a), (Wang et al., 2017), (Kim et al., 2017), (Liu et al., 2017), (Yi et al., 2017), (Mejjati et al., 2018), (Tang et al., 2019) (left to right) on oranges to apples task. Leftmost column shows the input, rightmost column shows results from our method.

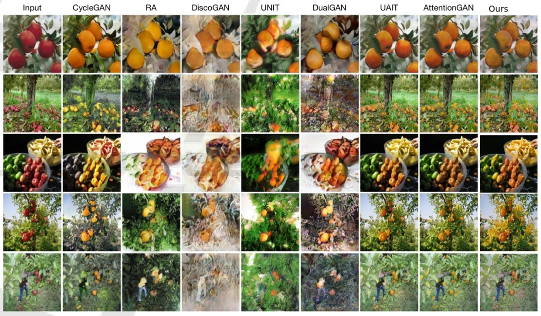

Figure 6: Comparison of test images generated by different methods (Zhu et al., 2017a), (Wang et al., 2017), (Kim et al., 2017), (Liu et al., 2017), (Yi et al., 2017), (Mejjati et al., 2018), (Tang et al., 2019) (left to right) on apples to oranges task. Leftmost column shows the input, rightmost column shows results from our method.

6 RESULTS AND DISCUSSION

The results from Table 1 suggest that there is a slight gain in the performance by using dropout with the Resnet network architecture for the horse → zebra task, while the same is more or less not true for apples → oranges task. We hypothesize that this observation could be due to the simplicity of the apples → oranges task while the former task is more complex. The results from Table 3 and Table 2 suggest that our method is at par to existing image translation methods where some related methods have an upper hand due to the underlying attention mechanisms.

On comparing the qualitative results for *zebras* → *horses* in Fig.3, in the second row we can notice that the text color and the background are preserved only in the translated images from UAIT (Mejjati et al., 2018), AttentionGAN (Tang et al., 2019) and our method. Also, our method has comparable results to (Mejjati et al., 2018) and (Tang et al., 2019), which uses attention mechanisms. While CycleGAN does a great job in translating all zebra images to horse images, the background color and tint are affected in some of the images and is severe than ours.

On comparing the qualitative results for *horses* → *zebras* in Fig. 4, we notice that residual attention based method (Wang et al., 2017) generates convincing translation results, while there is a green tint in all the translated images which makes it unfavourable. Similarly, UNIT (Liu et al., 2017) also has artifacts in the background that makes the appearance just acceptable. Our method falls behind UAIT (Mejjati et al., 2018) and AttentionGAN results (Tang et al., 2019) but appears better than CycleGAN results, which has undesirable background tint in many images. Note that the translation quality drastically dropped for DualGAN (Yi et al., 2017), with slightly better results from DiscoGAN (Kim et al., 2017).

On comparing the qualitative results for *oranges* → *apples* in Fig. 5, we notice that our translation results are at par with UAIT (Mejjati et al., 2018) and AttentionGAN (Tang et al., 2019), which uses attention mechanisms. CycleGAN also follows our results except that it fails in some of the images with unwanted background translations. DualGAN, DiscoGAN, UNIT and residual attention (Wang et al., 2017) fails on a task much easier than the *horses* ↔ *zebras* task.

On comparing the qualitative results for *apples* → *oranges* in Fig. 6, we notice that our method consistently keeps the quality upto the mark of attention guided methods (Mejjati et al., 2018), (Tang et al., 2019). While CycleGAN is able to translate convincingly, it is affected by a strong tint in some of the images. The translation results from UNIT and residual attention method appear similar to that of DualGAN and DiscoGAN, despite the use of attention mechanisms.

Table 4: Ablation study results on baseline, variant$_1$, variant$_2$, variant$_3$ and variant$_4$ (from left to right) evaluated using FID scores. The study included training on 128x128 size images and testing on 256x256 size images.

FID	D_{shared}	$D_{\text{shared}}1$	No Image pool	No stage-1	No Stage-2
Horse	**207.93** ± **6.26**	218.74 ± 4.69	216.10 ± 7.659	221.04 ± 5.005	224.02 ± 6.056
Zebra	**92.91** ± **6.58**	100.90 ± 7.495	136.63 ± 10.444	139.77 ± 10.643	119.39 ± 7.025

Ablation Study. We perform an ablation study to isolate the effects and understand the effectiveness of various components of our method using FID and KID metric over *horse* ↔ *zebra* task comparing the baseline to different variants. We consider the original shared discriminator setup, D_{shared} to be the baseline for comparing the variants. The quantitative results of the ablation study are available in Table 4 and 5 for images trained on 128x128 size images and tested on 256x256 size images and in Table 6 and 7 for images both trained and tested on 256x256 size images.

First, we modify the objective of the shared discriminator in a variant D_{shared_1} to see if all six components are really necessary. Out of the six components of the shared discriminator objective, four of them involves random source or target domain images conditioned on either a' or b'. It may seem logical to remove two random image components of one translation task or the other to make the shared discriminator objective compact, since they differ only in the conditioned part, i.e. a' or b'. To verify that, we deal with each domain only once and as target domain in the variant D_{shared_1}. The source domain images conditioned on the input images are not subjected to the shared discriminator and avoided in the objective assuming that the same domain images as target domain and labels will suffice. In other words, for the image translation task $A \rightarrow B$, we consider only random target domain images b from B with *true* labels conditioned on the input images a' to the shared discriminator. Analogously, for the image translation task $B \rightarrow A$, we consider only random target domain images a from A, conditioned on the input images b' with *false* labels to the shared discriminator. The generator's goal and objectives are unaltered in this variant. The results in Table 4, 5, 6 and 7 from

Table 5: Ablation study results on baseline, variant$_1$, variant$_2$, variant$_3$ and variant$_4$ (from left to right) evaluated using KID scores. The study included training on 128x128 size images and testing on 256x256 size images.

KID	D_{shared}	$D_{\text{shared}}1$	No Image pool	No stage-1	No Stage-2
Horse	**0.065** ± **0.003**	0.084 ± 0.002	0.088 ± 0.002	0.085 ± 0.002	0.107 ± 0.002
Zebra	**0.036** ± **0.002**	0.047 ± 0.003	0.063 ± 0.003	0.067 ± 0.003	0.050 ± 0.002

Table 6: Ablation study results on baseline, variant$_1$, variant$_2$, variant$_3$ and variant$_4$ (from left to right) trained and tested on 256x256 sizes and evaluated using FID scores.

FID	D_{shared}	$D_{\text{shared}}1$	No Image pool	No stage-1	No Stage-2
Horse	**212.81** ± **4.835**	221.66 ± 6.185	213.64 ± 4.357	216.28 ± 4.884	217.67 ± 6.864
Zebra	**92.72** ± **9.915**	148.95 ± 4.470	96.16 ± 5.251	118.63 ± 9.380	113.30 ± 11.212

Table 7: Ablation study results on baseline, variant$_1$, variant$_2$, variant$_3$ and variant$_4$ (from left to right) trained and tested on 256x256 sizes and evaluated using KID scores.

KID	D_{shared}	$D_{\text{shared}}1$	No Image pool	No stage-1	No Stage-2
Horse	**0.069** ± **0.002**	0.090 ± 0.002	0.070 ± 0.002	0.072 ± 0.002	0.077 ± 0.002
Zebra	**0.030** ± **0.002**	0.076± 0.004	0.036 ± 0.003	0.047 ± 0.003	0.045 ± 0.003

the ablation study for D_{shared_1}(or $variant_1$) indicate that irrespective of the image sizes used for training, the performance of the shared discriminator setup drops on removing the components of D_{shared}'s objectives. Both FID and KID values have gone higher for D_{shared_1}, which is not desirable for good image translation results.

The second variant that we consider is a shared discriminator setup without the image pool, i.e. the translated images are not reused as inputs to any of the generators. While the results in Table 6 and 7 suggest that using image pool doesn't improve the performance of our method, the results in Table 4 and 5 suggest that there is considerable drop in performance when the image pool is not used while training on smaller images and testing the model on larger images. We hypothesize that the generators become more robust when trained with additional translated images with the help of image pools.

The third variant that we consider is a shared discriminator setup without the training stage-1, i.e. the translated images are pushed to the image pool of the inverse translation task, throughout the training process. Similarly, the fourth variant that we consider is a shared discriminator setup without the training stage-2, i.e. the translated images are pushed back to the image pool of the same translation task throughout the training process. The results in Table 6 suggest that FID values are not really affected for $variant_3$ and $variant_4$, while Table 7 suggests that the KID values increase (or performance drops) for $variant_3$ and $variant_4$ when either training stage-1 or stage-2 is used throughout the training process. Similarly, the results from Table 4 and 5 for $variant_3$ and $variant_4$ indicates that the performance drops on using only one of the training stages. We hypothesize that simply reusing translated images with an image pool doesn't improve the performance and can result in a drop in performance.

7 SUMMARY

In this paper, we propose a framework for image transfiguration, a cross-domain image-to-image translation task improving the efficacy using a shared discriminator in an unsupervised setting. We also introduce a novel application of image pools to keep the generators more robust in the process. The qualitative and quantitative results, using metrics like FID and KID, suggest that our method, even without using masks or attention mechanisms, is at par with attention-based methods. For particular tasks, where the source domain shares similar semantics with the target domain, our method performs better than previous methods. Also, we observe that metrics like KID and FID are insufficient to evaluate the quality of translated images. They are also vulnerable to adversarial noise and hallucinated features, hampering a fair comparison of image translation methods. Future work could use attention mechanisms to further improve the results and better comparison metrics that correlate to human perception.

REFERENCES

Bińkowski, M., Sutherland, D. J., Arbel, M., and Gretton, A. (2018). Demystifying MMD GANs. *arXiv e-prints*, page arXiv:1801.01401.

Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., and Krishnan, D. (2016). Unsupervised pixel-level domain adaptation with generative adversarial networks. *CoRR*, abs/1612.05424.

Choi, Y., Choi, M., Kim, M., Ha, J. W., Kim, S., and Choo, J. (2018). StarGAN: Unified Generative Adversarial Networks for Multi-domain Image-to-Image Translation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 8789–8797.

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., and Bharath, A. A. (2018). Generative adversarial networks: An overview. *IEEE Signal Processing Magazine*, 35(1):53–65.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in NIPS 27*, pages 2672–2680. Curran Associates, Inc.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Klambauer, G., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a nash equilibrium. *CoRR*, abs/1706.08500.

Huang, X., Liu, M., Belongie, S. J., and Kautz, J. (2018). Multimodal unsupervised image-to-image translation. *CoRR*, abs/1804.04732.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *CVPR*.

Jabbar, A., Li, X., and Omar, B. (2020). A survey on generative adversarial networks: Variants, applications, and training.

Johnson, J., Alahi, A., and Fei-Fei, L. (2016). Perceptual Losses for Real-Time Style Transfer and Super-Resolution.

Kastaniotis, D., Ntinou, I., Tsourounis, D., Economou, G., and Fotopoulos, S. (2018). Attention-aware generative adversarial networks (ata-gans). *CoRR*, abs/1802.09070.

Kim, T., Cha, M., Kim, H., Lee, J. K., and Kim, J. (2017). Learning to discover cross-domain relations with generative adversarial networks. *CoRR*, abs/1703.05192.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *CoRR*, abs/1312.6114.

Kurach, K., Lucic, M., Zhai, X., Michalski, M., and Gelly, S. (2018). The GAN landscape: Losses, architectures, regularization, and normalization. *CoRR*, abs/1807.04720.

Ledig, C., Theis, L., Huszar, F., Caballero, J., Aitken, A. P., Tejani, A., Totz, J., Wang, Z., and Shi, W. (2016). Photo-realistic single image super-resolution using a generative adversarial network. *CoRR*, abs/1609.04802.

Liang, X., Zhang, H., and Xing, E. P. (2017). Generative semantic manipulation with contrasting GAN. *CoRR*, abs/1708.00315.

Liu, M., Breuel, T., and Kautz, J. (2017). Unsupervised image-to-image translation networks. *CoRR*, abs/1703.00848.

Liu, M., Huang, X., Mallya, A., Karras, T., Aila, T., Lehtinen, J., and Kautz, J. (2019). Few-shot unsupervised image-to-image translation. *CoRR*, abs/1905.01723.

Liu, M. and Tuzel, O. (2016). Coupled generative adversarial networks. *CoRR*, abs/1606.07536.

Ma, L., Jia, X., Sun, Q., Schiele, B., Tuytelaars, T., and Gool, L. V. (2017). Pose guided person image generation. *CoRR*, abs/1705.09368.

Mejjati, Y. A., Richardt, C., Tompkin, J., Cosker, D., and Kim, K. I. (2018). Unsupervised attention-guided image to image translation. *CoRR*, abs/1806.02311.

Mirza, M. and Osindero, S. (2014). Conditional generative adversarial nets. *CoRR*, abs/1411.1784.

Park, T., Liu, M.-Y., Wang, T.-C., and Zhu, J.-Y. (2019). Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Qian, R., Tan, R. T., Yang, W., Su, J., and Liu, J. (2017). Attentive generative adversarial network for raindrop removal from a single image.

Radford, A., Metz, L., and Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597.

Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., and Webb, R. (2016). Learning from simulated and unsupervised images through adversarial training. *CoRR*, abs/1612.07828.

Taigman, Y., Polyak, A., and Wolf, L. (2016). Unsupervised cross-domain image generation. *CoRR*, abs/1611.02200.

Talreja, V., Taherkhani, F., Valenti, M. C., and Nasrabadi, N. M. (2019). Attribute-guided coupled gan for cross-resolution face recognition.

Tang, H., Liu, H., Xu, D., Torr, P. H. S., and Sebe, N. (2019). AttentionGAN: Unpaired Image-to-Image Translation using Attention-Guided Generative Adversarial Networks. *arXiv e-prints*, page arXiv:1911.11897.

Tang, H., Xu, D., Wang, W., Yan, Y., and Sebe, N. (2019). Dual Generator Generative Adversarial Networks for Multi-domain Image-to-Image Translation. volume 11361 LNCS, pages 3–21. Springer Verlag.

Tung, H. F., Harley, A. W., Seto, W., and Fragkiadaki, K. (2017). Adversarial inverse graphics networks: Learning 2d-to-3d lifting and image-to-image translation from unpaired supervision. *CoRR*, abs/1705.11166.

Vondrick, C., Pirsiavash, H., and Torralba, A. (2016). Generating videos with scene dynamics. *CoRR*, abs/1609.02612.

Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., and Tang, X. (2017). Residual attention network for image classification. *CoRR*, abs/1704.06904.

Wang, T.-C., Liu, M.-Y., Tao, A., Liu, G., Kautz, J., and Catanzaro, B. (2019). Few-shot Video-to-Video Synthesis. *arXiv e-prints*, page arXiv:1910.12713.

Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Liu, G., Tao, A., Kautz, J., and Catanzaro, B. (2018a). Video-to-video synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Wang, T.-C., Liu, M.-Y., Zhu, J.-Y., Tao, A., Kautz, J., and Catanzaro, B. (2018b). High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*.

Wang, Z., Chen, J., and Hoi, S. C. H. (2019). Deep Learning for Image Super-resolution: A Survey.

Wang, Z., She, Q., and Ward, T. E. (2020). Generative adversarial networks in computer vision: A survey and taxonomy.

Wu, H., Zheng, S., Zhang, J., and Huang, K. (2017). GP-GAN: Towards Realistic High-Resolution Image Blending.

Yang, C., Kim, T., Wang, R., Peng, H., and Kuo, C. J. (2018). Show, attend and translate: Unsupervised image translation with self-regularization and attention. *CoRR*, abs/1806.06195.

Yi, Z., Zhang, H., Tan, P., and Gong, M. (2017). Dualgan: Unsupervised dual learning for image-to-image translation. *CoRR*.

Zhang, H., Goodfellow, I., Metaxas, D., and Odena, A. (2018). Self-attention generative adversarial networks.

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017a). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV 2017*.

Zhu, J.-Y., Zhang, R., Pathak, D., Darrell, T., Efros, A. A., Wang, O., and Shechtman, E. (2017b). Toward multimodal image-to-image translation. In *Advances in NIPS*.