Unaligned Image-to-Image Translation by Learning to Reweight

Shaoan Xie[1], Mingming Gong[2], Yanwu Xu[3], and Kun Zhang[1]

[1]Carnegie Mellon University, [2] The University of Melbourne, [3] University of Pittsburgh

Abstract

Unsupervised image-to-image translation aims at learning the mapping from the source to target domain without using paired images for training. An essential yet restrictive assumption for unsupervised image translation is that the two domains are aligned, e.g., for the selfie2anime task, the anime (selfie) domain must contain only anime (selfie) face images that can be translated to some images in the other domain. Collecting aligned domains can be laborious and needs lots of attention. In this paper, we consider the task of image translation between two unaligned domains, which may arise for various possible reasons. To solve this problem, we propose to select images based on importance reweighting and develop a method to learn the weights and perform translation simultaneously and automatically. We compare the proposed method with state-of-the-art image translation approaches and present qualitative and quantitative results on different tasks with unaligned domains. Extensive empirical evidence demonstrates the usefulness of the proposed problem formulation and the superiority of our method.

1. Introduction

In recent years, Image-to-Image (I2I) translation has been achieving remarkable success in transferring complex appearance changes across domains [61, 34]. In addition, many related tasks could also be formulated as I2I problems such as image super-resolution [57, 11] and domain adaptation [23, 42].

In supervised image translation, we are given paired data from source and target domains. Pix2pix [28] applies conditional Generative Adversarial Network [17, 40] to map the the source images to the target domain while enforcing a L1 distance loss between translated images and target images. Pix2pix can generate a sharp target image with sufficient paired training data. However, paired data are very difficult to collect or even do not exist (e.g., Van Gogh's painting to real photos). In the absence of paired data, unsupervised I2I translation methods have achieved impressive performance by combining GAN with proper constraints, such as cycle consistency [61] and shared latent space assumption [34].

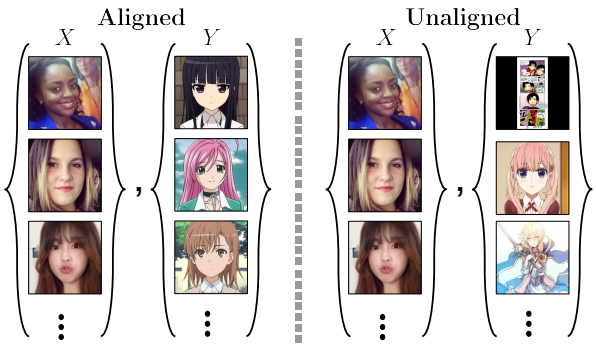


Figure 1: Example of aligned and unaligned domains. Left: selfie images as domain X and anime face images as domain Y. Images in two domains are carefully selected and processed. Right: many unwanted anime images may appear in the domain Y for many possible reasons, e.g., lack of human supervision.

An essential assumption of unsupervised image translation is then that the domains used for training are *aligned*, which means that each image in one domain can be translated to some image in the other domain in a meaningful manner; in other words, there is some underlying relationship between the domains [61]. For example, each of the two domains in the selfie2anime task include only female face images (Figure 1, left) of a similar style.

However, collecting images for two domains which are guaranteed to be aligned needs a lot of attention. For instance, to collect the anime domain, Kim et al. [29] first constructed an initial dataset consisting of 69,926 anime character images. Then they applied pre-trained anime face detector to extract 27,073 face images and then manually selected 3500 female face images as the training set. To collect the animal face dataset, Liu et al. [35] manually labeled bounding boxes of 10,000 carnivorous animal faces in the images and selected images with high detection scores

from Imagenet [13].

To save efforts, one may consider the setting with unaligned domains since they are much cheaper to obtain. For example, to obtain the anime face domain, we may also apply the anime face detector as Kim et al. did, and then just treat the detected results as images in the desired domain. Without any human supervision, the constructed domain may contain many unwanted anime images, e.g., anime body or even anime book images as shown in Figure 1 (right). These unaligned images may harm the image translation quality and can even cause the failure of some image translation methods (e.g., see Figure 4).

We therefore seek an algorithm that can learn to translate between *unaligned* domains where some images in either domain may be unrelated to the main task (Figure 1, right) and thus should not be considered for translation. For brevity, we denote these images as unaligned images. We further assume that there are unknown, aligned subsets $X_a \subseteq X$ and $Y_a \subseteq Y$, and our task is to discover such unknown subsets automatically and simultaneously learn the mapping between them. Inferring the unknown subsets X_a and Y_a seems to be challenging since we are given only two unaligned domains X and Y. To address this issue, we propose to reweight (or "select") each sample with importance β during the adversarial distribution matching process. Ideally, if β is almost 0, then the image is not in the aligned subset and hence not considered for translation.

Thus the problem boils down to learning appropriate importance weight for each sample for the purpose of sensible translation. To address the importance weight estimation problem, we analyze the causal generating process of images and hypothesize that *images in X_a and Y_a can be translated to the other domain faster than images in unaligned subsets since X_a and Y_a share the same content category*. Then we propose the reweighted adversarial loss which enables us to approximate the density ratios as well as performing image translation between two unknown aligned subsets X_a, Y_a. In addition, we also propose an effective sample size loss to avoid importance networks giving trivial solutions. We apply the proposed method to various image-to-image translation problems and the large improvements over strong baselines on unaligned dataset demonstrate the efficacy of our proposed translation method as well as the validity of our hypothesis. Code and data are available at https://github.com/Mid-Push/IrwGAN.

2. Related Works

Image Translation Contemporary image-to-image translation approaches leverage the strong power of Generative Adversarial Network (GAN) [17] to generate high-fidelity images. Paired image-to-image translation methods adopt reconstruction loss between the result and target to preserve the content of the input image [28]. In contrast, there is no paired data available in the task of unsupervised image translation. To address this issue, cycle consistency is proposed to reduce the number of possible mappings in the function space. It enforces a one-to-one mapping between source and target domains and is shown to achieve impressive visual performance [30, 61, 54]; however, the one-to-one correspondence may not be enough to preserve content and many methods are proposed to facilitate better image translation [47, 29, 39, 49]. Alternatively, shared latent space assumption [27, 33, 32, 34, 35] and relationship preservation [4, 60, 15, 44, 2] also demonstrated their efficacy in image translation. Recently multi-modal and multi-domain translation are gaining wide popularity [41, 37, 8, 27, 33, 32, 55, 12, 62, 1, 45, 43]. However, unlike the above prior works, we aim to learn the mapping with unaligned domains.

Importance Reweighting Importance reweighting is an important technique in various fields, including domain adaptation [26, 59, 18, 46, 56, 16, 53] and label-noise learning [36, 52, 58]. Their settings are very different from the problem we aim to solve. We are not given two sets of data points from which the density ratio is to be estimated. We use importance reweighting as a way to learn to select proper images for translation. Furthermore, in our task, we need to reweight samples in two domains simultaneously. This treatment, together with the property that aligned image subsets are easier to be translated to each other, helps achieve automated image selection and translation. Without this property, the problem to be solve might be ill-posed. There are also some importance reweighting applications on generative models [50, 48, 19, 14, 51, 9, 24, 22]. In this line of research, [51] reweights the fake samples of the generator by the exponential score of discriminator output to help discriminator training. [24] proposes to assign normalized discriminator score for samples to achieve a tighter lower bound for GAN following the importance reweighting Variational autoencoder [7]. [9] proposes to reweight each sample with the normalized ratio of discriminator output on a delayed copy of the generator to stabilize the training of GAN. These models [24, 51, 9] apply known statistics to reweight the samples for better GAN training. In contrast, our importance weight is unknown and our goal is to learn such importance weight.

3. Image Translation with Unaligned Domains

Given collected yet unaligned domains X and Y with training samples $\{x_i\}_{i=1}^N \in X$ and $\{y_j\}_{j=1}^M \in Y$, our goal is to translate proper images from one domain to the other domain while alleviating the detrimental effects brought by the unaligned images in two domains. We denote the aligned subsets of two domains by X_a and Y_a, respectively. On the other hand, $X_u = X \setminus X_a$ and $Y_u = Y \setminus Y_a$ correspond to the subsets that are not aligned in two domains.

Our model includes two mappings $G : X \rightarrow Y$ and $F : Y \rightarrow X$ and two discriminators D_Y and D_X. In addition, we introduce two importance weight networks β_X and β_Y , where β_X is the weight applied on $\{x_i\}$ and β_Y the weight on $\{y_j\}$. Naturally, our objective includes four types of terms: a reweighted adversarial loss to learn importance weights and perform image translation, an effective sample size loss to control how many images are selected (in a soft manner) by importance weights, a reweighted cycle consistency loss to keep the one-to-one mapping between two domains and a reweighted identity loss to keep networks conservative. For brevity, we refer to our method as Importance Reweighting Generative Adversarial Network (`IrwGAN`).

3.1. Connection between Aligned and Unaligned Translation

If the given domains X and Y are aligned (which is not the case in our setting), to transfer images from X to Y, one may apply least squares adversarial loss [38] directly to match the distributions between P_Y and $P_{G(X)}$ with the mapping function $G : X \rightarrow Y$ and its discriminator D_Y by solving:

$$\begin{aligned}&\min_G \max_{D_Y} \quad \mathcal{L}_{\text{GAN}}(X, Y) \\ &=\mathbb{E}_{x\sim P_X}[(1 - D_Y(G(x)))^2] + \mathbb{E}_{y\sim P_Y}[D_Y(y)^2].\end{aligned} \tag{1}$$

However, X and Y are unaligned. Matching the unaligned images can harm the translation performance and we only want to match the distributions of unknown aligned subsets, In other words, we would like to optimize $\mathcal{L}_{\text{GAN}}(X_a, Y_a)$. We can observe that

$$\begin{aligned}&\mathcal{L}_{\text{GAN}}(X_a, Y_a) \\ &=\mathbb{E}_{x\sim P_{X_a}}[(1 - D_Y(G(x)))^2] + \mathbb{E}_{y\sim P_{Y_a}}[D_Y(y)^2] \\ &=\int_x P_X(x)\frac{P_{X_a}(x)}{P_X(x)}[(1 - D_Y(G(x)))^2]dx+ \\ &\qquad\qquad \int_y P_Y(y)\frac{P_{Y_a}(y)}{P_Y(y)}[D_Y(y)^2]dy \\ &=\mathbb{E}_{x\sim P_X}\frac{\boldsymbol{P_{X_a}(x)}}{\boldsymbol{P_X(x)}}[(1 - D_Y(G(x)))^2]+ \\ &\qquad\qquad \mathbb{E}_{y\sim P_Y}\frac{\boldsymbol{P_{Y_a}(y)}}{\boldsymbol{P_Y(y)}}[D_Y(y)^2].\end{aligned} \tag{2}$$

It shows that even though the given domains X, Y are unaligned, if we reweight each sample $x \in X$ and $y \in Y$ with the corresponding density ratio $\frac{P_{X_a}(x)}{P_X(x)}$ and $\frac{P_{Y_a}(y)}{P_Y(y)}$, respectively, we are actually optimizing $\mathcal{L}_{\text{GAN}}(X_a, Y_a)$.

However, the density ratios are unknown. In light of this observation, we apply two networks β_X, β_Y and use their outputs to reweight each sample. If we are able to learn $\beta_X(x) \approx \frac{P_{X_a}(x)}{P_X(x)}, \beta_Y(y) \approx \frac{P_{Y_a}(y)}{P_Y(y)}$, optimizing the generators G, F and discriminators D_Y, D_X is equivalent to performing image translation between unknown aligned subsets X_a, Y_a.

3.2. Learning to Select Images and Translate

As shown above, we aim to find the density ratios with our importance networks β_X and β_Y, without having access to aligned subsets. However, although we can restrict the outputs of β_X and β_Y to be valid density ratios, there can still exist many unwanted solutions. For example, images not in the aligned subsets may be translated to each other with neural networks of high capacity and long enough training; then the estimated importance weights could be just one for all images, failing to select the correct aligned subset. Hence, we need to formulate and exploit proper constraints on the property of aligned subsets to find meaningful solutions of β_X, β_Y and achieve successful translation.

To this end, we formulate the following *quicker translation hypothesis*: *images in X_a and Y_a can be translated to the other domain faster than images in X_u and Y_u*. With this hypothesis, the images in the aligned set will first be selected to match in the two domains, while the images in the unaligned subsets will have a relatively large adversarial loss and hence achieve very small importance weights.

More specifically, we propose the following *reweighted adversarial loss* to estimate the density ratios as well as matching the distributions:

$$\begin{aligned}&\min_{G,\beta_X} \max_{D_Y} \quad \mathcal{L}_{\text{GAN}}(X, Y) \\ &= \mathbb{E}_{x\sim P_X}\beta_X(x)[(1 - D_Y(G(x)))^2]+ \\ &\qquad \mathbb{E}_{y\sim P_Y}\beta_Y(y)[D_Y(y)^2],\end{aligned} \tag{3}$$

where $\beta_X(x)$ and $\beta_Y(y)$ represent the importance weights assigned to each sample. We introduce a similar reweighted loss for the mapping function $F : Y \rightarrow X$, i.e., $\mathcal{L}_{\text{GAN}}(Y, X)$. Note that β_Y is not optimized in this loss function.

In each iteration, after we update the discriminator, we minimize the loss function $\mathcal{L}_{\text{GAN}}(X, Y)$, given in Equation (3), over β_X and G. Intuitively, according to our hypothesis, images $x_a \in X_a$ are translated to the domain Y faster, and thus the discriminator will assign a higher score to $G(x_a)$ than $G(x_u)$. Then $[(1 - D_Y(G(x_a)))^2] \leq [(1 - D_Y(G(x_u)))^2]$. If we assume that higher loss implies decreasing faster with the existing SGD optimization algorithms, then a consequence is that $\beta_X(x_u)$, compared to $\beta_X(x_a)$, will be decreased with a larger rate proportional to $[(1 - D_Y(G(x_u)))^2]$ as the optimization procedure (3) proceeds. $\beta_X(x_u)$ will then become smaller and finally are expected to be close to 0. In the experiment section 5.5, we visualize the learned importance weights and observe that the importance weights for unaligned images are close to 0.

3.3. Analysis of the Hypothesis

Although the quicker translation hypothesis is intuitive, one may wonder how sensible it is and how it helps in image

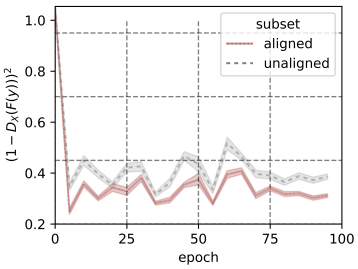


Figure 3: The trend of $(1-D_X(F(y)))^2$ for the task $A_U \rightarrow S$ where $X = S$ is the selfie face domain and $Y = A_U$ is the anime domain which contains anime faces (aligned subsets) and non-face anime images (unaligned subsets).

selection for translation. In this section, we will look into the generating process of images and provide an analysis of the hypothesis. Illustrative experiments are also provided to support our hypothesis.

Similar to the partial shared latent space assumption in [27, 32], we assume that *content category* C, *style* S and *random input* E (causally) generate the final image I, as illustrated in Figure 2. Naturally, images in aligned subsets X_a, Y_a are expected to share the same content category and differ only in the influence of style. Let us consider the same content category, denoted by C_1, which corresponds to the aligned subsets. Corresponding images are generated according to $I = F_{C_1,S}(E)$, where F is the generation function, S is the style indicator, and E is the random input. (E introduces the randomness in the images with the same content category and the same style, which is expected to remain the same during translation.)

The causal process is $C \rightarrow I$, and the two domains have different styles. The minimal change principle of the causal systems [25], as well as the quicker adaptation method for learning causality from data with changing distributions [5] suggests that $F_{C1,S2}$ is "close" to $F_{C1,S1}$; it is easy to adapt one to the other, given that the cause C_1 does not change. As a consequence, the translation function from style S_1 to S_2, which can be written as $I_Y = F_{C1,S2}(F^{-1}_{C1,S2}(I_X))$, is rather simple and easy to learn as well.

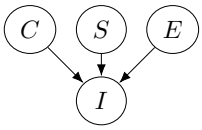


Figure 2: Causal generating process of images.

If we also change the content category from C_1 and C_2, in addition to the change in style, then the (unaligned) image subsets are generated by functions F_{C_1,S_1} and F_{C_2,S_2} in the source and target domains, respectively. The two functions may be greatly different because of the additional change in the content category. The (unnatural) translation function, which can be written as $F_{C_2,S_2}(F^{-1}_{C_1,S_2}(I_X))$, is expected to be more complex and difficult to learn than that corresponds to the same content category.

Now let us illustrate the hypothesis with real images. We use existing aligned face selfie2anime dataset as aligned subsets and we construct the unaligned subsets with non-face anime images. We apply CycleGAN [61] on the unaligned dataset (We replace the discriminator with our improved discriminator (see section 4)). As illustrated in Figure 3, the term $(1 - D_X(F(y)))^2$ is consistently larger for those unaligned images Y_u. An important consequence of our hypothesis is that $(1-D_X(F(y)))^2$ should be larger for unaligned images. The empirical evidence well aligns with this consequence. For more detail of this illustration, please refer to the supplementary material.

3.4. Effective Sample Size for Translation

Density ratio constraint. Previous analysis suggests that β_X and β_Y should approximate the density ratios. Thus we regularize our importance network outputs such that they are valid density ratios. Below we discuss the constraints on β_X, which also apply to β_Y. The first constraint is that it has to be non-negative and the second constraint is that the L_1 norm is fixed. The reason for the first constraint is obvious and as for the second constraint, we have $\mathbb{E}_{x\sim P_X}\frac{P_{X_a}(x)}{P_X(x)} = \int P_X(x)\frac{P_{X_a}(x)}{P_X(x)}dx = 1$. Thus we require the outputs of importance networks β_X, β_Y to have expectation 1. Empirically we need $\frac{1}{n}\sum_{i=1}^{n}\beta_X(x_i) = 1$ for a batch of images $\{x_1, ..., x_n\}$ where n is the batch size. A same constraint also applies to β_Y.

Effective sample size loss . The reweighted adversarial loss discards the unaligned images effectively through assigning them low importance weights. But there is no constraint on importance weights for aligned images. One may end up with the trivial case where networks assign low values to almost all images (including aligned images), that is, only few images are selected for translation. This is also observed in section 5.5. To address this issue, we propose the *effective sample size loss* to allow more aligned images being selected for translation:

$$\min_{\beta_X} \mathcal{L}_{\text{ESS}}(X) = ||\beta_X||_2. \tag{4}$$

A same loss $\mathcal{L}_{\text{ESS}}(Y)$ is introduced on the importance network β_Y. To understand why minimizing the above function will maximize the size of the paired subset for translation, one can see that under only the fixed L_1 norm constraint on β_X, the above term is minimized when $\beta_X(x) = 1$ for every image x in domain X, which means that all images are selected for translation. In contrast, if we only select one image for translation, i.e., $\beta_X(x_i) = n$ for one image x_i while $\beta_X(x_j) = 0$ for the remaining images $\{x_j\}$, $||\beta_X||_2$ reaches its maximum. Therefore, we can control the effective sample size by assigning different weights on the objective in Equation (4). Our effective sample size loss was inspired by [18] which proves that the effective sample size in domain adaptation may be defined as $n/||\beta_X||^2$

under some conditions. Therefore, we encourage networks to select more samples by penalizing the L_2 norm of the importance weight vectors.

3.5. Additional Regularizations

Adversarial loss in Equation (3) can help match the distributions between P_{Y_a} and $P_{G(X_a)}$. However, with a large enough capacity, a network can map the same set of input images to any random permutation of images in the target domain [61] and thus we need additional constraints to avoid it. To further regularize the mapping networks G and F, we also apply the *reweighted cycle consistency loss* to enforce a one-to-one mapping between the importance reweighted domain distributions:

$$\min_{G,F} \mathcal{L}_{\text{cyc}}(X,Y) = \mathbb{E}_{x\sim P_X}\beta_X(x)\|x - F(G(x))\|_1 + \mathbb{E}_{y\sim P_Y}\beta_Y(y)\|y - G(F(y))\|_1, \quad (5)$$

together with the *reweighted identity loss* to keep networks conservative [61]:

$$\min_{G,F} \mathcal{L}_{\text{idt}}(X,Y) = \mathbb{E}_{x\sim P_X}\beta_X(x)\|x - F(x)\|_1 + \mathbb{E}_{y\sim P_Y}\beta_Y(y)\|y - G(y)\|_1. \quad (6)$$

3.6. Full Objective

Our full objective is

$$\begin{aligned}\min_{G,F}\max_{D_X,D_Y} \quad & \mathcal{L}_{\text{GAN}}(X,Y) \ + \mathcal{L}_{\text{GAN}}(Y,X) \\ & + \lambda_{\text{cyc}}\mathcal{L}_{\text{cyc}}(X,Y) + \lambda_{\text{idt}}\mathcal{L}_{\text{idt}}(X,Y), \\ \min_{\beta_X} \mathcal{L}_{\text{GAN}}(X,Y) + \lambda_{\text{ESS}}\mathcal{L}_{\text{ESS}}(Y), \\ \min_{\beta_Y} \mathcal{L}_{\text{GAN}}(Y,X) + \lambda_{\text{ESS}}\mathcal{L}_{\text{ESS}}(Y), \end{aligned} \quad (7)$$

where λ_{cyc}, λ_{idt}, and λ_{ESS} control the relative importance of different losses. In addition, importance weight vectors β_X and β_Y need to be non-negative and have fixed L_1 norm, as addressed by parameterization in section 4.

4. Implementation

Mapping and discriminator network architecture We adopt the generator architecture used in CycleGAN [61], which contains 9 residual blocks [20]. To capture the global structure and local region, we apply two discriminators for each mapping direction; one consists of 3 downsampling convolutional layers and the other consists of 5 downsampling convolutional layers.

Importance weight network architecture The outputs have to satisfy the *non-negativity* and *fixed sum* constraints, and we address these two constraints by reparameterization: For the importance networks β_X and β_Y, we firstly downsample images to 64×64 in order to save memory. Then we apply 4 convolutional networks with kernel size 4, stride 2, padding 1. Then we append a fully connected network to the output and use a Softmax layer to normalize the outputs so that they sum to 1 and non-negative. Finally, we multiply the outputs by the batch size to make β satisfy the fixed sum constraint.

Training details For all the experiments, we set $\lambda_{\text{idt}} = \lambda_{\text{cyc}} = 10$ in Equation (7). We find that $\lambda_{\text{ESS}} = 1$ works well for all our experiments. We use the Adam solver [31] with the learning rate of 0.0001. We train networks from scratch and keep the same learning rate for the first 50 epochs and linearly decay the rate to zero over the next 50 epochs. We use 10,000 images in each epoch. Since we need to input a batch of images into the network at the same time while the resolution of input images is high, e.g., 256×256, we use the gradient accumulation trick to avoid GPU memory explosion. We set the batch size to 20 in all experiments.

5. Experiments

We first compare our approach against recent methods for unsupervised image translation on different datasets. We then present the evaluation results of learned importance weights β_X, β_Y. Finally, we investigate the effect of the proposed *effective sample size loss* by varying λ_{ESS}.

5.1. Dataset

For simplicity, we use abbreviations for the datasets: (S)elfie, (A)nime, (H)orse, (Z)ebra, (C)at, (D)og, danb(U)roo, (T)iger, tiger (B)eetle, (L)ion and s(E)a lion. P_Q denotes the domain after add images in Q to the original domain P.

$H_S \leftrightarrow Z_D$ and $C_H \leftrightarrow D_A$. Because most of existing image translation datasets are carefully built to be aligned, to evaluate our method, we first construct two unaligned datasets using three image translation datasets: horse2zebra [61] , selfie2anime [29] , and cat2dog [32] . For the constructed dataset $H_S \leftrightarrow Z_D$, the main task is horse2zebra; we add selfie domain to the horse domain and we add dog domain to the zebra domain. For $C_H \leftrightarrow D_A$, the main task is cat2dog, we add horse domain to the cat domain and add anime domain to the dog domain.

$S \leftrightarrow A_U$. To collect selfie2anime dataset, Kim et al. [29] used the pretrained face detector to collect anime faces. It highly depends on the accuracy of the face detector. Therefore, it is interesting to consider the case where there is no pretrained face detector or the detector accuracy is low. To this end, we add 2869 anime images from the Danbooru anime dataset [3] to the anime face domain. The Danbooru dataset covers anime face, body, book, and many related

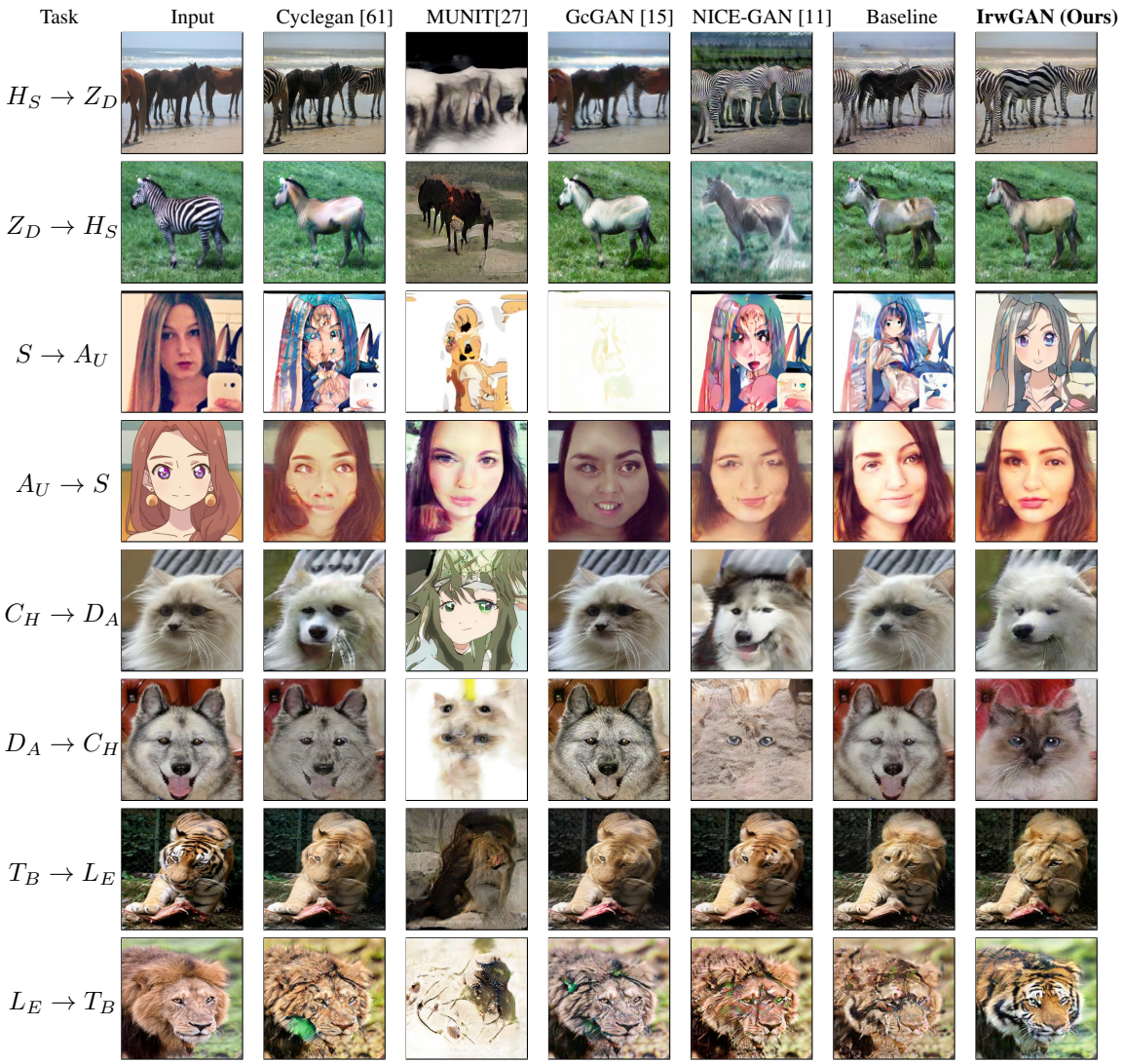


Figure 4: : Visual comparisons of results by different algorithms (see the top row) on different datasets (given on the left). Abbreviations: (S)elfie, (A)nime, (H)orse, (Z)ebra, (C)at, (D)og, danb(U)roo, (T)iger, tiger (B)eetle, (L)ion, s(E)a lion, le(O)pard. P_Q denotes the domain after add images in Q to the original domain P.

images. Since only one domain is unaligned, we only learn β_Y for this task.

$\boldsymbol{T_B \leftrightarrow L_E}$**.** We also consider a more realistic case where one uses search engines to obtain images: when searching for *lion*, we may get not only images related to lion but sea lion images. In light of this observation, we use tiger class (1300 images) and tiger beetle class (1300 images) in Imagenet [13] as T_B domain and lion class (1300 images) and sea lion class (1300 images) as L_E domain. We select 100 tiger images and 100 lion images as the test set.

5.2. Baselines and Metrics

We compare our method with CycleGAN [61], MUNIT [27], GcGAN [15], and NiceGAN [10]. Different from CycleGAN, we use one global and local discriminator for each translation mapping and adopt different learning rate and batch size. To exclude possible effects by these differences and fully examine the effects of our proposed method, we run our method with $\beta_X(x) = \beta_Y(y) = 1$ for all samples and we denote it as Baseline. For performance evaluation, we adopt the two commonly used metrics in image translation literature: the FID [21] and KID score [6]. They measure the distribution divergence between the generated images and target images.

5.3. Comparisons against Baselines

As can be seen in Figure 4, our method can produce good translation results with unaligned domains. In contrast, existing methods are unable to detect the unaligned

Table 1: The FID and KID (×100) for different algorithms. Lower is better. Abbreviations:(S)elfie, (A)nime, (H)orse, (Z)ebra, (C)at, (D)og, danb(U)roo, (T)iger, tiger (B)eetle, s(E)a lion. P_Q denotes the unaligned domain after add images in Q to the original domain P.

Method	$H_S \to Z_D$		$S \to A_U$		$C_H \to D_A$		$T_B \to L_E$	
	FID ↓	KID ↓	FID ↓	KID ↓	FID ↓	KID ↓	FID ↓	KID ↓
CycleGAN [61]	87.28	2.74	100.43	**1.98**	68.97	2.03	112.41	6.95
MUNIT [27]	287.79	22.19	180.95	8.10	132.21	5.77	335.52	25.35
GcGAN [15]	174.38	11.32	267.73	20.92	73.59	2.47	110.76	6.48
NICE-GAN [10]	166.61	3.79	124.11	4.51	229.00	18.80	147.32	11.99
Baseline	106.84	4.01	123.35	4.09	64.79	**1.89**	97.82	3.05
IrwGAN	**79.40**	**1.83**	**93.75**	2.58	**60.97**	2.07	**84.91**	**2.34**

Method	$Z_D \to H_S$		$A_U \to S$		$D_A \to C_H$		$L_E \to T_B$	
	FID ↓	KID ↓	FID ↓	KID ↓	FID ↓	KID ↓	FID ↓	KID ↓
CycleGAN [61]	151.94	4.63	124.46	2.29	96.94	3.43	101.32	4.75
MUNIT [27]	245.97	10.59	127.14	3.66	174.32	7.11	304.80	26.33
GcGAN [15]	161.75	**3.45**	133.58	3.77	153.83	8.71	130.91	8.83
NICE-GAN [10]	166.54	3.52	128.44	2.45	194.96	11.42	135.52	6.98
Baseline	162.32	3.73	**115.39**	2.24	61.28	2.06	112.77	4.82
IrwGAN	**142.98**	3.74	119.86	**2.07**	**53.46**	**1.84**	**77.47**	**2.44**

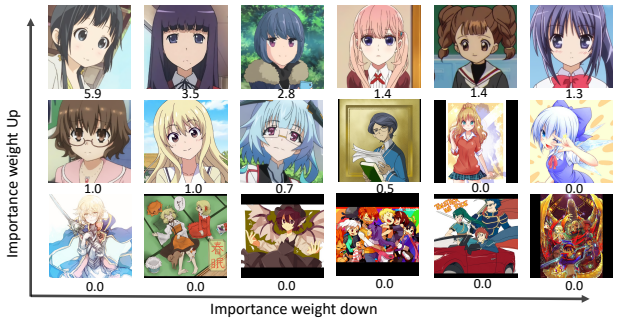


Figure 5: Examples of the learned importance weights for domain A_U in the task $S \leftrightarrow A_U$.

images and tend to generate unrelated images. For example, in the task $S \to A_U$, given in the third row, the main task is selfie2anime but most of existing methods are heavily influenced by the Danbooru anime images and hence produce messy results. In particular, Baseline translates a female selfie image to an anime character image rather than the wanted anime face image.

Table 1 reports the FID and KID values of different image translation tasks. Our method outperforms these strong baselines on most datasets. The clear improvement of IrwGAN compared to Baseline method suggests that the importance reweighting scheme plays a crucial role in obtaining good image translation results when domains are unaligned.

5.4. Analysis of Importance Reweighting

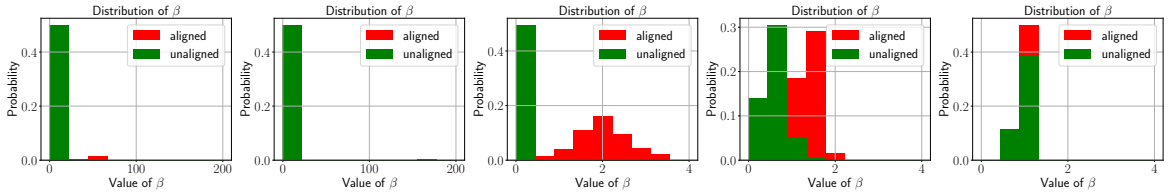
Figure 5 visualizes the learned weights for domain A_U in the task $S \to A_U$. As we can see, our method IrwGAN is able to distinguish unaligned images from images in the aligned subsets. Many unwanted anime images, e.g., the full body anime character image in the third row, are assigned very low values of the importance weight. As a consequence, these unwanted images would not affect our image translation process and thus we obtain the best results compared to those by other methods.

Table 2: Precision, recall and accuracy score for the learned β in different domains. Baseline denotes our method in which we assign 1 to each sample; its recall is always 1.00.

Domain	Method	Precision	Recall	Accuracy
H_S	Baseline	0.55	**1.00**	0.55
	IrwGAN	**1.00**	0.93	**0.96**
Z_D	Baseline	0.58	**1.00**	0.58
	IrwGAN	**1.00**	0.64	**0.79**
A_U	Baseline	0.50	**1.00**	0.50
	IrwGAN	**0.99**	0.97	**0.98**
C_H	Baseline	0.45	**1.00**	0.45
	IrwGAN	**0.99**	0.97	**0.98**
D_A	Baseline	0.50	**1.00**	0.50
	IrwGAN	**1.00**	**1.00**	**1.00**
T_B	Baseline	0.50	**1.00**	0.50
	IrwGAN	**0.80**	0.89	**0.83**
L_E	Baseline	0.50	**1.00**	0.50
	IrwGAN	**0.78**	0.85	**0.81**

Table 2 shows the performance the learned β_X and β_Y on different unaligned domains. For unaligned domain P_Q, we set labels for images in domain P as 1 and 0 for domain Q. Since our learned β is continuous, for evaluation purposes, we consider its prediction as 1 if it is above a predefined threshold 0.5. Our method IrwGAN outperforms the Basline method by a large margin in terms of precision and accuracy. Note that Baseline achieves a perfect recall score because its prediction is always 1 and thus the false negative is 0.

Importance reweighting helps to recover the aligned subsets from two unaligned domains. It would be interest-



$\lambda_{\text{ESS}} = 0.0$ $\lambda_{\text{ESS}} = 0.1$ $\lambda_{\text{ESS}} = 1.0$ $\lambda_{\text{ESS}} = 3.0$ $\lambda_{\text{ESS}} = 10.0$

Figure 6: Distributions of importance weights β_X with different λ_{ESS} for domain C_H in the task $C_H \leftrightarrow D_A$. We use red to denote importance weights for images in aligned subset C and green for images in unaligned subset H.

Table 3: Results of $S \rightarrow A_U$ under different ratios of unaligned to aligned samples.

Ratio	FID ↓	Ratio	FID ↓
0 %	95.25	100 %	93.73
10 %	95.58	150 %	92.74
30 %	90.54	200 %	91.25
50 %	93.05	300 %	95.44

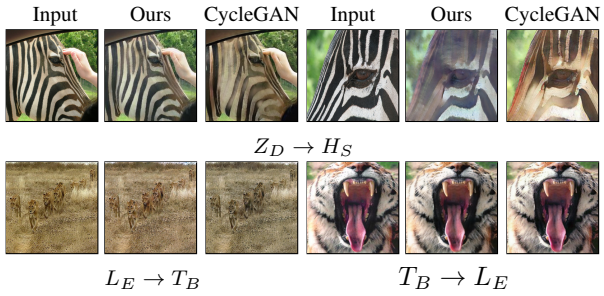


Figure 7: Some failure cases of IrwGAN.

ing to test how many unaligned samples that our method can handle. We first use the existing dataset selfie2anime as the aligned subsets and then add N anime images from the Danbooru dataset to the anime domain. We set N to be $10, 30, 50, 100, 150, 200, 300\%$ of the number of images (3400) in the original anime domain. Results in Table 3 show that our method is capable of handling the unaligned translation problem under different levels of unaligned samples. We also provide visualizations of learned weights in the supplementary material.

5.5. The Effective Sample Size Weight λ_{ESS}

λ_{ESS} is designed to control the effective sample size. Figure 6 shows the distributions of the estimated importance weights β_X for domain C_H with different λ_{ESS} in the task $C_H \leftrightarrow D_A$. As we can see, if λ_{ESS} is set to 0 or very low, the vector is very sparse, which means our method can only select a few images from the whole domain. As we increase the value of λ_{ESS}, the importance weights of more images in unknown aligned subsets are becoming larger while the importance weights of unaligned images are still very small. If we set λ_{ESS} to a very large value, e.g., 10, the importance weights of all images concentrate around 1.0, which is very close to Baseline that assigns 1 to each sample.

6. Conclusion and Discussion of Limitations

In this paper, we proposed a novel, more realistic setting for image-to-image translation, in which the two domains are not aligned and hence one has to select suitable images for meaningful translation. To show that the formulated problem is not only more practical, but also solvable, we developed an importance reweighting-based learning method to automatically select the images and perform translation simultaneously. Our empirical results suggest that it achieves large improvements over existing methods. It is worth noting that our method relies on the assumption that aligned images are easier to be translated to each other. We observe that this hypothesis is generally supported on real images, although it might be violated for some complex images and specific network structure. In other words, it might also be hard to translate some images in the aligned subsets to the other domain. This violation may result in low importance weights on those samples and they will consequently be discarded during training.

Figure 7 shows some failure cases of our method. In the first row, we want to translate zebra images to horse images. However, there are few horse head images in the horse domain, which makes it difficult to translate zebra head images to the horse domain. As a consequence, our model will assign low importance to these images and they are discarded during the training. Our method and CycleGAN both failed on this task. CycleGAN made little changes to the input image while ours method outputs almost identical image to the input. Similar phenomena happened in the second row. Resolving this issue may require some weak supervision or additional information for representation learning and we leave it as future work. In addition, the domain gap between aligned and unaligned subsets may also be an important factor of the performance. We plan to explore dataset with more diverse domain gaps in the future work.

Acknowledgements We would like to acknowledge the support by the United States Air Force under Contract No. FA8650-17-C-7715, by National Institutes of Health under Contract No. R01HL159805, and by a grant from Apple. The United States Air Force or National Institutes of Health is not responsible for the views reported in this article. Mingming Gong was supported by Australian Research Council Project DE210101624.

References

[1] Amjad Almahairi, Sai Rajeswar, Alessandro Sordoni, Philip Bachman, and Aaron Courville. Augmented cyclegan: Learning many-to-many mappings from unpaired data. *arXiv preprint arXiv:1802.10151*, 2018.

[2] Matthew Amodio and Smita Krishnaswamy. Travelgan: Image-to-image translation by transformation vector learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8983–8992, 2019.

[3] Anonymous, Danbooru community, and Gwern Branwen. Danbooru2019: A large-scale crowdsourced and tagged anime illustration dataset. `https://www.gwern.net/Danbooru2019`, January 2020. Accessed: DATE.

[4] Sagie Benaim and Lior Wolf. One-sided unsupervised domain mapping. In *Advances in neural information processing systems*, pages 752–762, 2017.

[5] Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Rosemary Ke, Sébastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher Pal. A meta-transfer objective for learning to disentangle causal mechanisms. *arXiv preprint arXiv:1901.10912*, 2019.

[6] Mikołaj Bińkowski, Dougal J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

[7] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.

[8] Jiezhang Cao, Langyuan Mo, Yifan Zhang, Kui Jia, Chunhua Shen, and Mingkui Tan. Multi-marginal wasserstein gan. In *Advances in Neural Information Processing Systems*, pages 1774–1784, 2019.

[9] Tong Che, Yanran Li, Ruixiang Zhang, R Devon Hjelm, Wenjie Li, Yangqiu Song, and Yoshua Bengio. Maximum-likelihood augmented discrete generative adversarial networks. *arXiv preprint arXiv:1702.07983*, 2017.

[10] Runfa Chen, Wenbing Huang, Binghui Huang, Fuchun Sun, and Bin Fang. Reusing discriminators for encoding: Towards unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8168–8177, 2020.

[11] Shuaijun Chen, Zhen Han, Enyan Dai, Xu Jia, Ziluan Liu, Liu Xing, Xueyi Zou, Chunjing Xu, Jianzhuang Liu, and Qi Tian. Unsupervised image super-resolution with an indirect supervised path. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 468–469, 2020.

[12] 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. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018.

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[14] Maurice Diesendruck, Ethan R Elenberg, Rajat Sen, Guy W Cole, Sanjay Shakkottai, and Sinead A Williamson. Importance weighted generative networks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 249–265. Springer, 2019.

[15] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, Kun Zhang, and Dacheng Tao. Geometry-consistent generative adversarial networks for one-sided unsupervised domain mapping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2427–2436, 2019.

[16] Mingming Gong, Kun Zhang, Tongliang Liu, Dacheng Tao, Clark Glymour, and Bernhard Schölkopf. Domain adaptation with conditional transferable components. In *International conference on machine learning*, pages 2839–2848, 2016.

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[18] Arthur Gretton, Alex Smola, Jiayuan Huang, Marcel Schmittfull, Karsten Borgwardt, and Bernhard Schölkopf. Covariate shift by kernel mean matching. *Dataset shift in machine learning*, 3(4):5, 2009.

[19] Aditya Grover, Jiaming Song, Ashish Kapoor, Kenneth Tran, Alekh Agarwal, Eric J Horvitz, and Stefano Ermon. Bias correction of learned generative models using likelihood-free importance weighting. In *Advances in Neural Information Processing Systems*, pages 11058–11070, 2019.

[20] 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.

[21] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, pages 6626–6637, 2017.

[22] R Devon Hjelm, Athul Paul Jacob, Tong Che, Adam Trischler, Kyunghyun Cho, and Yoshua Bengio. Boundary-seeking generative adversarial networks. *arXiv preprint arXiv:1702.08431*, 2017.

[23] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. PMLR, 2018.

[24] Zhiting Hu, Zichao Yang, Ruslan Salakhutdinov, and Eric P Xing. On unifying deep generative models. *arXiv preprint arXiv:1706.00550*, 2017.

[25] Biwei Huang, Kun Zhang, Jiji Zhang, Joseph Ramsey, Ruben Sanchez-Romero, Clark Glymour, and Bernhard Schölkopf. Causal discovery from heterogeneous/nonstationary data. *Journal of Machine Learning Research*, 21:1–53, 2020.

[26] Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex J Smola. Correcting sample selec-

tion bias by unlabeled data. In *Advances in neural information processing systems*, pages 601–608, 2007.

[27] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–189, 2018.

[28] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[29] Junho Kim, Minjae Kim, Hyeonwoo Kang, and Kwanghee Lee. U-gat-it: unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. *arXiv preprint arXiv:1907.10830*, 2019.

[30] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1857–1865. JMLR. org, 2017.

[31] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[32] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European conference on computer vision (ECCV)*, pages 35–51, 2018.

[33] 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.

[34] 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.

[35] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10551–10560, 2019.

[36] Tongliang Liu and Dacheng Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on pattern analysis and machine intelligence*, 38(3):447–461, 2015.

[37] Liqian Ma, Xu Jia, Stamatios Georgoulis, Tinne Tuytelaars, and Luc Van Gool. Exemplar guided unsupervised image-to-image translation with semantic consistency. *arXiv preprint arXiv:1805.11145*, 2018.

[38] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2794–2802, 2017.

[39] 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.

[40] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[41] Sangwoo Mo, Minsu Cho, and Jinwoo Shin. Instagan: Instance-aware image-to-image translation. *arXiv preprint arXiv:1812.10889*, 2018.

[42] Zak Murez, Soheil Kolouri, David Kriegman, Ravi Ramamoorthi, and Kyungnam Kim. Image to image translation for domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4500–4509, 2018.

[43] Ori Nizan and Ayellet Tal. Breaking the cycle-colleagues are all you need. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7860–7869, 2020.

[44] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. *arXiv preprint arXiv:2007.15651*, 2020.

[45] Zhiqiang Shen, Mingyang Huang, Jianping Shi, Xiangyang Xue, and Thomas S Huang. Towards instance-level image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3683–3692, 2019.

[46] Masashi Sugiyama, Taiji Suzuki, Shinichi Nakajima, Hisashi Kashima, Paul von Bünau, and Motoaki Kawanabe. Direct importance estimation for covariate shift adaptation. *Annals of the Institute of Statistical Mathematics*, 60(4):699–746, 2008.

[47] Hao Tang, Dan Xu, Nicu Sebe, and Yan Yan. Attention-guided generative adversarial networks for unsupervised image-to-image translation. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019.

[48] Chenyang Tao, Liqun Chen, Ricardo Henao, Jianfeng Feng, and Lawrence Carin Duke. Chi-square generative adversarial network. In *International conference on machine learning*, pages 4887–4896, 2018.

[49] Wayne Wu, Kaidi Cao, Cheng Li, Chen Qian, and Chen Change Loy. Transgaga: Geometry-aware unsupervised image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8012–8021, 2019.

[50] Yifan Wu, Tianshu Ren, and Lidan Mu. Importance reweighting using adversarial-collaborative training. In *NIPS 2016 Workshop*, 2016.

[51] Yue Wu, Pan Zhou, Andrew Gordon Wilson, Eric P Xing, and Zhiting Hu. Improving gan training with probability ratio clipping and sample reweighting. *arXiv preprint arXiv:2006.06900*, 2020.

[52] Xiaobo Xia, Tongliang Liu, Nannan Wang, Bo Han, Chen Gong, Gang Niu, and Masashi Sugiyama. Are anchor points really indispensable in label-noise learning? In *Advances in Neural Information Processing Systems*, pages 6838–6849, 2019.

[53] Hongliang Yan, Yukang Ding, Peihua Li, Qilong Wang, Yong Xu, and Wangmeng Zuo. Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE Conference*

on Computer Vision and Pattern Recognition, pages 2272–2281, 2017.

[54] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE international conference on computer vision*, pages 2849–2857, 2017.

[55] Xiaoming Yu, Yuanqi Chen, Shan Liu, Thomas Li, and Ge Li. Multi-mapping image-to-image translation via learning disentanglement. In *Advances in Neural Information Processing Systems*, pages 2990–2999, 2019.

[56] Xiyu Yu, Tongliang Liu, Mingming Gong, Kun Zhang, Kayhan Batmanghelich, and Dacheng Tao. Label-noise robust domain adaptation. ICML, 2020.

[57] Yuan Yuan, Siyuan Liu, Jiawei Zhang, Yongbing Zhang, Chao Dong, and Liang Lin. Unsupervised image super-resolution using cycle-in-cycle generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 701–710, 2018.

[58] Kun Zhang, Mingming Gong, Petar Stojanov, Biwei Huang, QINGSONG LIU, and Clark Glymour. Domain adaptation as a problem of inference on graphical models. *Advances in Neural Information Processing Systems*, 33, 2020.

[59] Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. Domain adaptation under target and conditional shift. In *International Conference on Machine Learning*, pages 819–827, 2013.

[60] Rui Zhang, Tomas Pfister, and Jia Li. Harmonic unpaired image-to-image translation. *arXiv preprint arXiv:1902.09727*, 2019.

[61] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[62] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Advances in neural information processing systems*, pages 465–476, 2017.