Truly Unsupervised Image-to-Image Translation with Contrastive Representation Learning

Zhiwei Hong[1], Jianxing Feng[2], and Tao Jiang[1,3] ✉

[1] Tsinghua University, Beijing 100084, China
hzw17@mails.tsinghua.edu.cn
[2] Haohua Technology Co., Ltd, Shanghai, China
[3] University of California, Riverside, CA 92521, USA
jiang@cs.ucr.edu

Abstract. Image-to-image translation is a classic image generation task that attempts to translate an image from the source domain to an analogous image in the target domain. Recent advances in deep generative networks have shown remarkable capabilities in translating images among different domains. Most of these models either require pixel-level (with paired input and output images) or domain-level (with image domain labels) supervision to help the translation task. However, there are practical situations where the required supervisory information is difficult to collect and one would need to perform truly unsupervised image translation on a large number of images without paired image information or domain labels. In this paper, we present a truly unsupervised image-to-image translation model that performs the image translation task without any extra supervision. The crux of our model is an embedding network that extracts the domain and style information of the input style (or reference) image with contrastive representation learning and serves the translation module that actually carries out the translation task. The embedding network and the translation module can be integrated together for training and benefit from each other. Extensive experimental evaluation has been performed on various datasets concerning both cross-domain and multi-domain translation. The results demonstrate that our model outperforms the best truly unsupervised image-to-image translation model in the literature. In addition, our model can be easily adapted to take advantage of available domain labels to achieve a performance comparable to the best supervised image translation methods when all domain labels are known or a superior performance when only some domain labels are known.

1 Introduction

Image-to-image translation (I2I) is a classic image generation task [25] that attempts to translate an image from the source domain to an analogous image in the target domain while preserving the content representations. I2I has attracted extensive attention nowadays due to its wide range of applications in many computer vision and image processing tasks such as image synthesis, image style transfer, human pose estimation, *etc*. Thanks to the rapid development

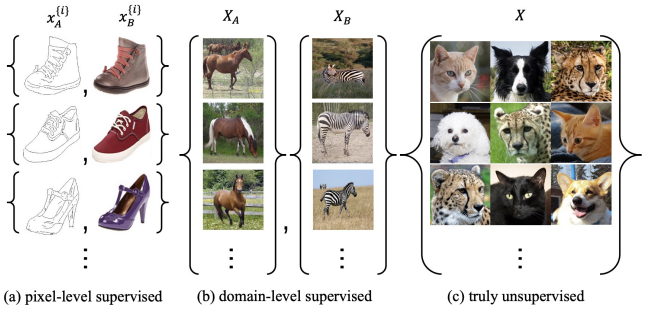

Fig. 1. Different levels of supervision. (**a**) Pixel-level supervision consists of training examples $\{x_A^{(i)}, x_B^{(i)}\}_{i=1}^N$, where $x_A^{(i)}$ and $x_B^{(i)}$ have matching content. (**b**) Domain-level supervision consists of a source set $X_A = \{x_A^{(i)}\}_{i=1}^N$ of images from domain A and a target set $X_B = \{x_B^{(j)}\}_{j=1}^N$ of images from domain B, with no information provided as to which $x_B^{(j)}$ matches each $x_A^{(i)}$. The multi-domain translation scenario is similar. (**c**) A truly unsupervised instance simply consists of a set of images X where neither matching image pairs nor domain information are provided.

of deep neural networks especially generative adversarial networks, image-to-image translation has achieved remarkable progress in the past few years. Isola *et al*. [13] first used a conditional Generative Adversarial Network (cGAN) [23] to perform image-to-image translation with pixel-level supervised input-output image pairs. But the applicability of this method seems to be restricted in many real situations, including image synthesis and style transfer, where such matching image pairs are not available. Therefor, some unsupervised image-to-image translation models such as CycleGAN [35] and UNIT [19] have been proposed to deal with image translation between two domains (*i.e.*, cross-domain image translation) without pixel-level supervision. In addition to cross-domain image translation, many multi-domain image translation models [12, 5, 20, 28, 18, 31] have been developed in the last few years. Though these models are generally called `unsupervised` in contrast to the pixel-level supervised methods [13, 27], they are actually not truly unsupervised, since they implicitly assume that the domain labels of the training images are given *a priori*. However, this assumption may be hard to satisfy in practice when the number of domains and samples increases. In particular, when we are given a large number of images from unknown sources (FFHQ [15]), it might be expensive and difficult to figure out the domain of each image, especially because some of the domain boundaries may be vague. So, the truly unsupervised image-to-image translation problem (where neither pixel-level paired images nor domain labels are available) has been introduced by Kyungjune *et al*. [2] recently. As described in [2], it can help reduce the effort of data annotation for model training and provide robustness against noisy labels produced by a manual labeling process. More importantly, it may also serve as a strong baseline for developing semi-supervised image translation models. Here, we are given two images x_A (the source or content image) and x_B (the reference or style image) and our goal is to translate x_A to an analogous image in the

same domain as x_B while preserving its original content. See Fig. 1 for a more detailed illustration of three levels of supervision in image translation. Though a solution to this problem was proposed in [2], it suffers from some serious issues such as the content loss issue discussed in COCO-FUNIT [28]. Since there is no supervision, it is difficult to control what parts of the input style image should be incorporated or transferred into the input content image. Ideally, the transferred information should only include style, such as fur texture and color in animal images. In reality, other types of information such as the pose of objects often get in as well. Hence, how to construct a proper embedding of the reference image is a critical step in truly unsupervised image translation. COCO-FUNIT [28] tried to reduce the style embedding (or code) variance of different input image crops to tackle this problem. More specifically, it utilizes the content embedding of the input content image to normalize the style embedding of the input style image, which helps to improve the translation performance but cannot completely eliminate the confusions between content and style representations.

In this paper, we extend the work in [2] and present a general **C**ontrastive representation learning based truly **UN**supervised **I**mage-to-image **T**ranslation model (**CUNIT**). Our overall method proceeds in three steps. First, we cluster images into (pseudo) domains to create pseudo domain labels for each image. Second, we extract the unique style embedding of the input style image. Finally, we learn to translate images between pseudo domains with the guide of style embeddings and pseudo domain labels. The first two steps are realized by using a style embedding network with two branches (or modules) that output the pseudo domain label and style embedding respectively. Here, pseudo domain labels are generated by a differentiable clustering method based on mutual information maximization [14]. To create the style embedding, a Siamese network architecture [4, 7], which is flexible with batch size and does not require negative samples, is adopted to tackle the content-loss problem discussed in COCO-FUNIT [28]. We try to ensure that the style embedding of an input image is close to those of its augmented versions (*e.g.*, images obtained by RandomResizedCrop [3]) but far from those of other images by using a normalized L_2 distance loss. The clustering and style embedding modules share a common encoder in the style embedding network so both can benefit from each other. To realize the last step of our method, a cGAN is adopted to perform reference-guided image translation. After integrating the style embedding network and cGAN together, our model is able to separate image domains and perform image translation smoothly under a truly unsupervised setting.

Extensive experimental evaluation has been performed on various datasets concerning both cross-domain and multi-domain scenarios. The results demonstrate that our model outperforms the best truly unsupervised I2I model in the literature and is comparable or even superior to the supervised I2I models when the domain labels are fully or partially provided. The major contributions of our work include: **(1)** We extend the work in [2] and present a general model for truly unsupervised image-to-image translation without requiring any explicit supervision (at neither the pixel-level nor domain-level). **(2)** We adopt a new

contrastive representation learning architecture to control the style embedding so as to help deal with the style code variance problem [28] and extract better style features. We also introduce a new reconstruction loss to better preserve content features. The superior performance of our model in extensive experiments over the state-of-the-art (SOTA) image-to-image translation methods demonstrate the effectiveness of both above techniques. **(3)** Our model could easily be adapted to take advantage of available domain labels to perform comparably to the best supervised image translation methods when all domain labels are known or significantly better when only some domain labels are known.

2 Related Work

Generative adversarial networks. Image generation and synthesis have been widely investigated in recent years. Different from auto-encoder architectures like VAE [17], generative adversarial networks (GANs) [6] play a zero-sum game and is composed of two parts: a generator G and a discriminator D. The generator G is trained to generate samples that are closed to real data from a random variable and D is trained to distinguish whether a sample is generated by G or from real data. Mehdi and Simon proposed conditional GANs [23] (cGANs) to generate data based on a particular condition. To address the stability issues in GANs, Wasserstein-GAN (WGAN) [1] was proposed to optimize an approximation of the Wasserstein distance. To further improve the vanishing and exploding gradient problems of WGAN, Gulrajani *et al*. [8] proposed WGAN-GP that uses gradient penalty instead of the weight clipping to deal with the Lipschitz constraint in WGAN.

Pixel-level supervised I2I. Isola *et al*. first proposed Pix2Pix [13] that utilizes a cGAN to do the image translation based on pixel-level supervised input-output image pairs. Following this seminal work, a sequence of I2I models have shown remarkable performance. For example, Wang *et al*. proposed pix2pixHD [32] to learn a mapping that converts a semantic image to a high-resolution photo-realistic image. Park *et al*. proposed SPADE [27] to further improve pix2pixHD on handling diverse input labels and delivering better output quality.

Domain-level supervised and truly unsupervised I2I. Apart from the above pixel-level supervised I2I, many unsupervised I2I methods have been introduced in the past few years. These so-called `unsupervised` methods do not need matching image pairs but still explicitly require the image domain information. Here, we call them domain-level supervised methods as opposed to our truly unsupervised setting. Zhu *et al*. proposed CycleGAN [35] to deal with cross-domain I2I with a cycle consistency loss. UNIT [19] tries to learn a one-to-one mapping between two visual domains based on a shared latent space assumption. MUNIT [12] further learns a many-to-many mapping between two visual domains. In MSGAN [21], a simple yet effective regularization term was proposed to address the mode collapse issue in cGANs that improved image diversity without loss of quality. Inspired by few-shot learning, Liu *et al*. proposed FUNIT [20] to learn a style-guided image translation model that can generate

translations in unseen domains. COCO-FUNIT [28] further improved FUNIT with a content-conditioned style encoding scheme for style code computation. Note that these methods all require image domain labels for training. Wang *et al*. [33] tried to utilize the noise-tolerant pseudo labeling scheme to reduce the labeling cost at the training process. Recently, Kyungjune *et al*. introduced the first method TUNIT [2] for performing the truly unsupervised I2I task.

Contrastive representation learning and unsupervised clustering. Unsupervised representation learning aims to extract informative features for downstream tasks without any human supervision. Self-supervised learning represented by contrastive learning has demonstrated its effectiveness and remarkable performance in unsupervised representation learning recently. Different from generative-based models such as the auto-encoder, contrastive learning is a discriminative-based scheme whose core idea is to attract different augmented views of the same image (positive pairs) and repulse augmented views of different images (negative pairs). Based on contrastive learning, SimCLR [3] used very large batch sizes and MoCo [9] built a dynamic dictionary with a queue and a moving-averaged encoder to deal with the memory bank problem [34]. They both require high-quality negative samples to achieve a good performance. Interestingly, subsequent work BYOL [7] suggested that negative samples are not necessary for contrastive learning, and Simsiam [4] claimed that simple Siamese networks can learn informative representations without using negative sample pairs, large batches or momentum encoders. These results have become the new SOTA self-supervised visual representation learning methods. On the other hand, IIC [14] utilized mutual information maximization in an unsupervised manner so that the model clusters images while assigning the images to clusters evenly. In this paper, we integrate both the unsupervised clustering and representation learning methods to deal with the truly unsupervised I2I downstream task.

3 Method

3.1 Overview

Let X be a dataset consisting of images from $K(\geq 2)$ different domains. Suppose that K is unknown and for each image $x_i \in X$, its domain label y_i is also unknown. We use $\hat{K}$ to denote the estimated value of K and treat $\hat{K}$ as a hyper-parameter in training. The goal of (truly) unsupervised image-to-image translation model is to translate a 'content' image x_A from some domain A to an analogous image of some domain B as specified by a reference 'style' image x_B (*i.e.*, the domain that contains x_B), while preserving the content information of x_A. Our model consists of three components: a style embedding network, a conditional generator and a multi-task domain-specific discriminator (as shown in Fig. 2). The style embedding network is the key component that outputs the pseudo domain label y_B and style embedding s_B of the input reference image x_B. Then style embedding s_B is fed into the conditional generator as a 'condition' to guide the translation. The pseudo domain label y_B is fed into the domain-specific discriminator that forces the generator to generate an image with the style (*e.g.*

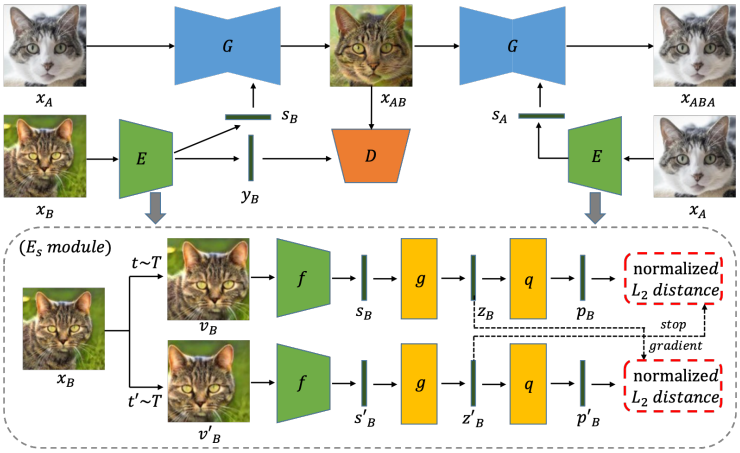

Fig. 2. An overview of the proposed CUNIT framework. The style embedding network E takes a reference image x_B as the input and estimates its pseudo domain label y_B and style embedding s_B. The pseudo domain label y_B is then used to train the domain-specific discriminator D. The style embedding s_B and the gradient feedback from D help the generator network G to translate the input content image x_A to the analogous image x_{AB} in the domain of x_B while preserving the content information of x_A.

fur texture and color in animal images) of x_B and content (*e.g.* object pose) of x_A. In general, we regard the features of an image that are not affected by various augmentation operations (such as cropping and affine transformations) as its styles while the others as its content. This is reflected in the choice of the loss function for training the contrastive style embedding module E_s (described below).

3.2 The style embedding network

The style embedding network E consists of two branches (or modules) E_y and E_s that output the pseudo domain label y and style embedding s respectively.

Unsupervised domain estimation E_y. The domain information, necessary for the subsequent domain-specific discriminator in our model, is unfortunately not available during training. To resolve this issue, we employ an unsupervised clustering approach to produce pseudo domain labels. Many methods have been proposed to deal with unsupervised image clustering with impressive performance in the past few years. Here, we adopt a differentiable clustering method called IIC [14] that maximizes the mutual information (MI) between feature vectors of two images. Given an image x, define $p = E_y(x)$ as the output of E_y, where p represents the probability vector of x over $\hat{K}$ domains. Similarly, we define x^+ and p^+ as the augmented versions of x and p. The mutual information between p and p^+ is thus $I(p, p^+) = H(p) - H(p|p^+)$. The value $I(p, p^+)$ reaches its optimum when the entropy $H(p)$ is maximized and the conditional entropy $H(p|p^+)$ is minimized. By maximizing the mutual information, E_y is

encouraged to distribute all images as evenly as possible over $\hat{K}$ domains while assigning paired images (x, x^+) to the same domain. The module E_y is trained via the following objective function:

$$\begin{aligned} &\mathcal{L}_{MI} = -I(\boldsymbol{p}, \boldsymbol{p}^+) = -I(\boldsymbol{P}) = -\sum_{i=1}^{\hat{K}}\sum_{j=1}^{\hat{K}} \boldsymbol{P}_{ij} \ln \frac{\boldsymbol{P}_{ij}}{\boldsymbol{P}_i \boldsymbol{P}_j}, \\ &s.t. \quad \boldsymbol{P} = \mathbb{E}_{x^+ \sim T(x)|x \sim p_{data}(x)}[E_y(x) \cdot E_y(x^+)^T] \end{aligned} \tag{1}$$

where T is a composition of random augmentations such as random cropping and affine transformations. $\boldsymbol{P}_i = \boldsymbol{P}(\boldsymbol{p} = i)$ denotes the $\hat{K}$-dimensional marginal probability vector, and $\boldsymbol{P}_{ij} = \boldsymbol{P}(\boldsymbol{p} = i, \boldsymbol{p}^+ = j)$ denotes the joint probability. (See [14] for more details of this objective function.) Here, the pseudo domain label $y = argmax(E_y(x))$ is generated as a one-hot vector to be fed to the domain-specific discriminator.

Contrastive style embedding E_s. To perform reference-guided image translation, a style embedding of the reference image is required for the generator. Inspired by [7, 4], we use contrastive learning based on a Siamese architecture to learn the style embedding. As mentioned in the last paragraph of section 2, such a model can learn informative representations without using negative sample pairs, large batches or momentum encoders. The Siamese architecture here is composed of an encoder f, a projector g, and a predictor q. During contrastive learning, the reference image x_B is randomly augmented by two transformations t and t' sampled from a transformation family T (as shown in Fig. 2) to generate two views v_B and v'_B. The transformation family includes widely used augmentations [3], such as RandomResizedCrop, RandomFlip, GaussianBlur, *etc*. The views v_B and v'_B are encoded by the encoder f to obtain the style embeddings s_B and s'_B. Then, the projector g and predictor q are applied to s_B or s'_B sequentially. For these two augmented views, denote the output of the projector g as $z_B \triangleq f(s_B)$ and $z'_B \triangleq f(s'_B)$ and the output of q as $p_B \triangleq q(z_B)$ and $p'_B \triangleq q(z'_B)$. We force p_B to be similar to z'_B and p'_B to be similar to z_B by minimizing the symmetric loss:

$$\mathcal{L}_{co}^E = \frac{1}{2}||\widetilde{p}_B - \widetilde{z'_B}||_2^2 + \frac{1}{2}||\widetilde{p'_B} - \widetilde{z}_B||_2^2 \tag{2}$$

where $\widetilde{\cdot} \triangleq \frac{\cdot}{||\cdot||_2}$ and $||\cdot||_2$ denotes the l_2 norm. Here, z_B and z'_B are detached from the computational graph before calculating the loss such that the gradient would not back-propagate through z_B and z'_B. In our experiments, the contrastive style embedding module significantly improves the quality of unsupervised image clustering compared to using only IIC [14] due to the shared encoder f between the two modules. A similar phenomenon has also been observed in [2].

3.3 The image translation module

As shown in Fig. 2, the image translation module is in fact a conditional generative adversarial network (cGAN) [23]. It takes both the original source domain

image x_A and the style embedding s_B of the reference image x_B as the input to generate x_{AB} that should have the same target domain label y_B as x_B. The style embedding s_B is fed to the decoding layers of G using a multi-scale AdaIN [11] technique. The discriminator D is a multi-task domain-specific discriminator [22] and it takes the pseudo domain label y_B as the input to guide the generator G to produce more realistic images. To train the entire image translation model, three loss functions are adopted. (a) The GAN loss is used to produce more realistic images in the target domain. (b) A style contrastive loss is used to further improve the quality of the generated images and prevent style corruption. (c) An image reconstruction loss is used to help the generated images preserve more content information (*i.e.*, domain-invariant features).

GAN loss. Given the content image x_A, reference image x_B, pseudo domain label y_B, and style embedding s_B, the GAN is trained with the following objective function:

$$\mathcal{L}_{adv} = \mathbb{E}_{x_B \sim p_{data}(x)}[\log D_{y_B}(x_B)] + \mathbb{E}_{x_A, x_B \sim p_{data}(x)}[\log(1 - D_{y_B}(G(x_A, s_B)))] \quad (3)$$

where $D_{y_B}(\cdot)$ denotes the logits from the domain-specific (y_B) discriminator. Note that there is no direct gradient backward propagation here from the discriminator D to style embedding network E because y_B is a one-hot vector only used to determine which head of the multi-task discriminator D to use.

Style contrastive loss. In order to prevent degenerate solutions where the generator ignores the given style embedding s_B and synthesizes a random image of domain B, we impose a style contrastive loss to the generator:

$$\mathcal{L}^G_{style} = \mathbb{E}_{x_A, x_B \sim p_{data}(x)}[-\log \frac{\exp(s_{AB} \cdot s_B)}{\sum_{i=0}^{N} \exp(s_{AB} \cdot s_i^- / \tau)}] \quad (4)$$

where $s_{AB} = E_s(x_{AB}) = E_s(G(x_A, s_B))$ denotes the style embedding of the translated image x_{AB} and s_i^- denotes the negative style embeddings (*i.e.*, style embeddings of other samples in the same mini-batch). This loss forces the generated image x_{AB} to have a dissimilar style to images other than the reference image x_B. It also prevents the encoder from mapping all images to the same style embedding.

Reconstruction loss. To better preserve the domain-invariant features (*i.e.*, the content information) of the content image x_A, an improved image reconstruction loss with a new term is introduced. The loss is composed of two parts, the `self-reconstruction` loss $\mathcal{L}_{self\_rec}$ and the `cross-reconstruction` loss $\mathcal{L}_{cross\_rec}$ (new term, similar to the work in [35]), as follows:

$$\mathcal{L}_{rec} = \mathbb{E}_{x_A \sim p_{data}(x)}[||x_A - G(x_A, s_A)||_1] + \mathbb{E}_{x_A, x_B \sim p_{data}(x)}[||x_A - G(G(x_A, s_B), s_A)||_1] \quad (5)$$

where $\mathcal{L}_{rec} = \mathcal{L}_{self\_rec} + \mathcal{L}_{cross\_rec}$ is intended to minimize the total l_1 distance between the source image x_A and its self-reconstructed image $G(x_A, s_A)$ and between x_A and its cross-reconstructed image $x_{ABA} = G(G(x_A, s_B), s_A)$. The reconstruction loss encourages G to preserve the domain-invariant information (*e.g.*, object pose).

3.4 Overall training

In our experiment, the unsupervised clustering module E_y and style embedding module E_s share a common encoder f in the style embedding network E so both can benefit from each other. The clustering module may obtain rich features acquired by contrastive representation learning in the style embedding module and improve its accuracy in generating pseudo domain labels. The style embedding module can also extract more domain-specific features and prevent the entire model from collapsing with the help of the clustering module. Once the embedding network E has been sufficiently trained, it can be further refined with the cGAN module jointly to perform image translation as follows. The generator G takes the style embedding extracted from input style image as a reference to translate the input content image to its analog that is expected to be in the same domain as the reference style image. With the adversarial loss feedback from the cGAN, the style embedding network E further improves its learned domain-separating features and extracts style embeddings with richer information that can help the generator G fool the domain-specific discriminator D. After integrating the style embedding network and the cGAN module together, our model is able to separate image domains and perform image-to-image translation successfully under a truly unsupervised setting.

The overall objective for above mentioned style embedding network E, generator G and discriminator D is given by

$$\begin{aligned} \mathcal{L}_D &= -\mathcal{L}_{adv} \\ \mathcal{L}_G &= \mathcal{L}_{adv} + \lambda^G_{style}\mathcal{L}^G_{style} + \lambda_{rec}\mathcal{L}_{rec} \\ \mathcal{L}_E &= \mathcal{L}_G + \lambda_{MI}\mathcal{L}_{MI} + \lambda_{co}\mathcal{L}^E_{co} \end{aligned} \tag{6}$$

where λ^G_{style}, λ_{rec}, λ_{MI}, and λ^E_{co} are weights for balancing different loss terms. More details about these parameters are given in the supplementary materials.

4 Experiments

In this section, we evaluate our model on both domain-labeled data and unlabeled data under different experimental settings (*i.e.*, cross-domain and multi-domain). The performance of our method is compared both quantitatively and qualitatively with that of the representative (and SOTA) published methods. The experiments are grouped according to (1) truly unsupervised settings and (2) semi-supervised settings when some (but not all) domain labels are available. We also perform ablation studies to assess the impact of the proposed objective functions, training strategy and choice of domain numbers.

Datasets. To evaluate performance on multi-domain image translation, we use the following three popular labeled datasets: AnimalFaces [20], Birds [30] and Flowers [24]. Following the strategies in [2], we select ten classes from each of the three dataset, referred to as *AnimalFaces-10*, *Birds-10* and *Flowers-10*. When these datasets are used in truly unsupervised image translation, their the

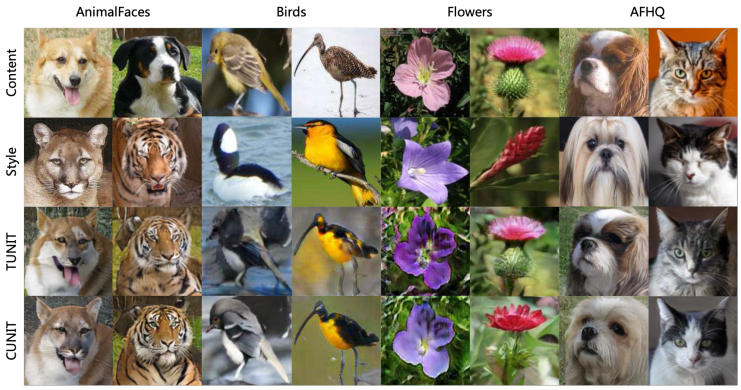

Fig. 3. Truly unsupervised image-to-image translation results on different datasets.

domain labels are simply masked. For cross-domain image translation evaluation, we use the dataset Summer2Winter and Dog2Cat from CycleGAN [35]. For data without domain labels, a high quality AFHQ [5] dataset is adopted. AFHQ involves roughly three groups of animals (cats, dogs and wild animals), where each group consists of diverse breeds/species with different styles but the exact domain labels are not provided.

Evaluation metrics and compared methods. We consider the following three metrics, *Inception Score (IS)* [29], the mean of class-wise *Fréchet Inception Distance (mFID)* [10] and *Translation Accuracy (Acc)* [20] in our experiments. The IS and mFID scores have been widely used in GAN-based image analyses to evaluate the generated image quality and diversity. The smaller an mFID score is, the better the performance is, which is the opposite for IS scores. The Acc score is used to evaluate whether a model is able to generate images of the same style as the target domain. It is a percent number between 0 and 100%. For experiments under truly unsupervised situation, we compare our model with TUNIT [2]. To the best of our knowledge, TUNIT is the only method that has been proposed in the literature to address the truly unsupervised I2I problem. For multi-domain supervised or semi-supervised image translation (where some but not all domain labels are given), we compare our model with COCO-FUNIT [28], SEMIT [33], Kim *et al*. [16] and TUNIT. Note that both TUNIT and our model can be easily adapted to take advantage of the available domain labels in their loss functions. We also compare our model with CycleGAN [35], UNIT [19], MSGAN [21], CUT [26] and COCO-FUNIT in cross-domain image translation task.

4.1 Truly unsupervised image-to-image translation

To verify that our proposed method is able to handle truly unsupervised image-to-image translation well, we evaluate our model on the datasets AnimalFaces-10, Birds-10, Flowers-10, and AFHQ. During the training, the domain labels of the first three datasets are masked as mentioned before. However, the labels are used later on for quantitative evaluation. Since AFHQ has three groups of images (cat,

Table 1. Quantitative evaluation of our model under truly unsupervised setting on different datasets with comparison to TUNIT [2]. The arrows indicate the directions of more desired values.

Dataset	Method	mFID ↓	IS ↑	Acc ↑
AnimalFaces-10	TUNIT	47.9	26.6	84.2
	CUNIT	45.2	28.9	88.3
Birds-10	TUNIT	82.3	73.8	62.2
	CUNIT	74.6	78.4	67.5
Flowers-10	TUNIT	67.3	48.7	65.8
	CUNIT	60.7	52.3	70.3

dog and wild), we train a separate model for each group. For all experiments on truly unsupervised image translation, we use TUNIT as the baseline and we set the number of clusters $\hat{K} = 10$ for all models as done in [2].

Fig. 3 and Table 1 show the visual results and quantitative evaluation of CUNIT and TUNIT on the four datasets. From the images in Fig. 3, we observe that CUNIT is able to capture more subtle style features (*e.g.* the fur textures of cats and dogs and the color information of birds and flowers) than TUNIT. More visual results can be found in the supplementary materials. Table 1 shows that CUNIT outperforms TUNIT by 5.6%, 8.6% and 5% in terms of mFID, IS and Acc, respectively, on AnimalFaces-10. On the datasets Birds-10 and Flowers-10, CUNIT outperformed TUNIT by at least 6% with respect to all evaluation metrics. Because the AFHQ dataset has no ground truth domain labels, we are not able to provide the quantitative scores on AFHQ. In summary, these experimental results suggest that our model outperforms TUNIT significantly in truly unsupervised image-to-image translation.

4.2 Domain-level supervised or semi-supervised image translation

Cross-domain (supervised) image translation. Recall that CycleGAN [35] adopts a cycle-consistency loss and UNIT [19] makes a shared-latent space assumption to learn a mapping between two visual domains, and both are representative models in cross-domain I2I. MSGAN [21] further improves the diversity of the generated images without loss of quality using a simple yet effective regularization term and has achieved impressive performance on cross-domain I2I. CUT [26] introduces contrastive learning to I2I successfully. COCO-FUNIT is the SOTA multi-domain image translation model that could also be applied in the cross-domain situation. We compare CUNIT with all above methods on the Summer2Winter and Dog2Cat datasets. Table 2 shows that CUNIT outperforms CycleGAN, UNIT, MSGAN, CUT, and COCO-FUNIT by 32% and 25%, 27% and 19%, 19% and 10%, 8% and 7%, and 4% and 3% in terms of mFID on the two datasets, respectively. Hence, CUNIT can generate images with more diversity and better quality than the SOTA cross-domain image translation methods. Some supportive visual results are given in supplementary Fig. S5.

Table 2. Quantitative evaluation (based on mFID) of our model with comparison to other methods in cross-domain image-to-image translation.

Method	CycleGAN [35]	UNIT [19]	MSGAN [21]	CUT [26]	COCO-FUNIT [28]	CUNIT (ours)
Summer2Winter	78.7	73.6	66.4	58.4	55.9	53.3
Dog2Cat	85.7	79.3	71.5	68.6	66.1	63.8

Table 3. Quantitative evaluation (based on mFID) in multi-domain semi-supervised image-to-image translation with partial domain labels available during training.

Method	AnimalFaces-10				Birds-10				Flowers-10			
	20%	40%	60%	80%	20%	40%	60%	80%	20%	40%	60%	80%
COCO-FUNIT [28]	104.8	85.9	75.3	59.6	129.5	108.7	98.3	82.9	128.6	105.2	87.5	70.4
SEMIT [33]	53.3	52.1	50.5	49.7	87.4	85.2	83.9	82.6	73.1	71.3	69.8	68.5
Kim *et al.* [16]	47.3	46.1	45.4	44.7	79.1	77.4	75.8	74.6	64.9	63.3	62.4	61.2
TUNIT [2]	46.2	46.2	45.9	45.6	80.5	80.1	79.6	79.5	65.7	65.4	64.9	65.0
CUNIT (ours)	44.9	44.1	44.3	43.9	73.8	73.2	73.5	72.9	59.8	59.6	59.1	58.9

Multi-domain semi-supervised image translation with partial domain labels. Since there are only two domains in cross-domain image translation, it is relatively easy to separate the domains and extract style embeddings for CUNIT. Now we evaluate the model's performance on multiple domains and compare it with the SOTA multi-domain image translation models COCO-FUNIT [28] (supervised), Kim *et al.* [16] and SEMIT [33] (semi-supervised), and TUNIT [2]. Recall that COCO-FUNIT is a domain-level fully supervised model while CUNIT and TUNIT are truly unsupervised. However, the latter two models can be easily modified to take advantage of the available labels as detailed below. In many practical situations, there may be a large number of images but only a few of them have domain labels. So, we also conduct a semi-supervised experiment as in [2] to test how these models perform when partial domain labels are provided.

The datasets AnimalFaces-10, Birds-10 and Flowers-10 are used in the experiment. Let X be a (whole) dataset. We separate X into the labeled part $X_{labeled}$ and unlabeled part $X_{unlabeled}$ with a ratio $\alpha = |X_{labeled}|/|X|$. Under this semi-supervised setting, we add an additional cross-entropy loss term to our model between the ground truth domain labels and pseudo domain labels estimated by the style embedding network on data $X_{labeled}$. The ground truth domain labels in $X_{labeled}$ are also used for training the domain-specific discriminator. A similar modification is also applied to TUNIT. In this experiment, we set the ratio α to 20%, 40%, 60%, and 80% as in [2]. The results are shown in Table 3. The performance of COCO-FUNIT significantly decays quickly when α decreases while CUNIT, TUNIT, Kim *et al.* [16] and SEMIT remain relatively stable. Although the latter four methods are more robust, CUNIT still outperforms the other three methods significantly in terms of mFID on all three datasets and under all ratios of α. This experiment shows that CUNIT can be easily adapted

Table 4. Ablation study on various components of CUNIT and training strategies using the AnimalFaces-10 dataset.

Configuration	mFID ↓	IS ↑	Acc ↑
CUNIT w\joint	45.2	28.9	88.3
CUNIT w\sequential	45.8	28.1	87.8
CUNIT w\o $\mathcal{L}_{co}$	48.5	26.5	83.4
CUNIT w\o $\mathcal{L}^{G}_{style}$	47.6	27.4	86.3
CUNIT w\o $\mathcal{L}_{rec}$	47.9	27.2	85.9
CUNIT w\o $\mathcal{L}_{cross\_rec}$	46.7	27.8	86.9

Table 5. Quantitative evaluation (based on mFID) of our model using different pseudo domain numbers $\hat{K}$ on AnimalFaces-10.

$\hat{K}$	1	4	7	10	13	16	20	30	50
mFID ↓	91.4	62.3	51.4	45.2	46.6	47.5	48.5	50.7	54.3

to the semi-supervised image translation scenario when some domain labels are available with a performance superior to the existing methods. Some supportive visual results are given in supplementary Fig. S6.

4.3 Some analyses of the proposed model

In this section, we analyze the impact of our proposed objective functions, training strategy and choice of domain numbers on the performance of CUNIT.

Ablation study. In Table 4, we ablate various components of our model and measure their impact on performance in truly unsupervised image-to-image translation on the AnimalFaces-10 dataset. We observe that joint training of the style embedding network and image translation network has a better performance than training the two models sequentially. The loss $\mathcal{L}_{co}$ is the most important one among the three loss terms ($\mathcal{L}_{co}$, $\mathcal{L}^{G}_{style}$ and $\mathcal{L}_{rec}$), which improved mFID by 3.3(6.8%), IS by 2.4(9%) and Acc by 4.9%. We also observe that the mFID score of CUNIT without using the cross-reconstruction loss term $\mathcal{L}_{cross\_rec}$ in Equation 5 was 46.7 on the AnimalFaces-10 dataset (vs 45.2 with the term), which clearly demonstrates the effectiveness of this new term. The results indicate that CUNIT has benefited a lot from the jointly trained style embedding network based on contrastive representation learning.

Sensitivity to $\hat{K}$. The hyper-parameter $\hat{K}$ (the number of pseudo domains used in training) may influence how CUNIT clusters the images and hence its performance in truly unsupervised I2I. We test CUNIT on the AnimalFaces-10 dataset with different values of $\hat{K}$ and summarize its performance in Table 5. As expected, the model achieves the best performance in terms of mFID when $\hat{K}$ is the same as the real domain numbers K ($K = 10$ in AnimalFaces-10). Moreover, when $\hat{K}$ slightly larger than K, the model still performed reasonably well. The (simple) experiment suggests that CUNIT has a relatively robust performance as

long as estimated $\hat{K}$ is near or slightly larger than the true K. More discussions concerning both K and $\hat{K}$ can be found in supplementary section 2.

Limitation of CUNIT. The above results demonstrate that our model CUNIT performs very well in reference-guided image translation under truly unsupervised, semi-supervised or even supervised settings when the number of domains is not very large. It would be interesting to know if this advantage of CUNIT remains true when the number of domains is very large. Supplementary Table S1 shows a comparison of CUNIT, TUNIT and COCO-FUNIT on the AnimalFaces dataset with various values of K, where CUNIT and TUNIT are trained without domain labels and COCO-FUNIT is trained with full labels. On AnimalFaces-10, CUNIT achieves a comparable performance as COCO-FUNIT (45.2 vs 44.8 in mFID) and outperforms TUNIT (47.9) by 5.6%. However, on AnimalFaces-149, the performance of both CUNIT and TUNIT drop significantly (106.9 and 106.3, respectively) compared with COCO-FUNIT (92.4). Because mFID measures the difference in feature distributions between the generated images and training images, the results suggest that, as K increases, it is very hard for the unsupervised methods to infer domain labels for images consistent with the true domain labels. Hence, CUNIT is more suitable in image translation applications where the domain numbers are not that large. More detailed analysis on how the number of domains K affects the performance of our model can be found in supplementary subsection 2.2 and Table S1.

5 Conclusion

Most of the existing I2I methods either require pixel-level or domain-level supervision to help the translation task. In this paper, we present a truly unsupervised I2I model, CUNIT, to perform image translation without requiring any supervision. The model consists of a style embedding network that extracts the domain and style information of the input style image with contrastive representation learning and an image translation module based on cGANs that actually carries out the reference-guided image translation. The embedding network and translation module are integrated together for training and benefit from each other, which enables CUNIT to successfully separate image domains and perform translation between these domains. Extensive experimental evaluation has been performed on various datasets concerning both cross-domain and multi-domain image translation. The results demonstrate that our model outperforms the best truly unsupervised I2I model in the literature (TUNIT). In addition, our model can be easily adapted to take advantage of the available domain labels to achieve a performance comparable to the best supervised image translation methods when all domain labels are known or a superior performance when only some (but not all) domain labels are provided. Therefor, we believe that CUNIT has great potentials in many practical image translation applications.

Acknowledgments: This work was supported in part by the National Key Research and Development Program of China grant 2018YFC0910404.

References

1. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: International Conference on Machine Learning. pp. 214–223 (2017)
2. Baek, K., Choi, Y., Uh, Y., Yoo, J., Shim, H.: Rethinking the truly unsupervised image-to-image translation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14154–14163 (2021)
3. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International conference on machine learning. pp. 1597–1607. PMLR (2020)
4. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15750–15758 (2021)
5. Choi, Y., Uh, Y., Yoo, J., Ha, J.W.: Stargan v2: Diverse image synthesis for multiple domains. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8188–8197 (2020)
6. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672–2680 (2014)
7. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.A., Guo, Z.D., Azar, M.G., et al.: Bootstrap your own latent: A new approach to self-supervised learning. arXiv preprint arXiv:2006.07733 (2020)
8. Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., Courville, A.C.: Improved training of wasserstein gans. In: Advances in Neural Information Processing Systems. pp. 5767–5777 (2017)
9. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9729–9738 (2020)
10. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
11. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1501–1510 (2017)
12. Huang, X., Liu, M.Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: Proceedings of the European conference on computer vision (ECCV). pp. 172–189 (2018)
13. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)
14. Ji, X., Henriques, J.F., Vedaldi, A.: Invariant information clustering for unsupervised image classification and segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9865–9874 (2019)
15. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4401–4410 (2019)
16. Kim, K., Park, S., Jeon, E., Kim, T., Kim, D.: A style-aware discriminator for controllable image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18239–18248 (2022)

17. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
18. Lin, J., Xia, Y., Liu, S., Zhao, S., Chen, Z.: Zstgan: An adversarial approach for unsupervised zero-shot image-to-image translation. Neurocomputing **461**, 327–335 (2021)
19. Liu, M.Y., Breuel, T., Kautz, J.: Unsupervised image-to-image translation networks. In: Advances in neural information processing systems. pp. 700–708 (2017)
20. Liu, M.Y., Huang, X., Mallya, A., Karras, T., Aila, T., Lehtinen, J., Kautz, J.: Few-shot unsupervised image-to-image translation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10551–10560 (2019)
21. Mao, Q., Lee, H.Y., Tseng, H.Y., Ma, S., Yang, M.H.: Mode seeking generative adversarial networks for diverse image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1429–1437 (2019)
22. Mescheder, L., Geiger, A., Nowozin, S.: Which training methods for gans do actually converge? In: International conference on machine learning. pp. 3481–3490. PMLR (2018)
23. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
24. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing. pp. 722–729. IEEE (2008)
25. Pang, Y., Lin, J., Qin, T., Chen, Z.: Image-to-image translation: Methods and applications. arXiv preprint arXiv:2101.08629 (2021)
26. Park, T., Efros, A.A., Zhang, R., Zhu, J.Y.: Contrastive learning for unpaired image-to-image translation. In: European conference on computer vision. pp. 319–345. Springer (2020)
27. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2337–2346 (2019)
28. Saito, K., Saenko, K., Liu, M.Y.: Coco-funit: Few-shot unsupervised image translation with a content conditioned style encoder. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16. pp. 382–398. Springer (2020)
29. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training gans. Advances in neural information processing systems **29**, 2234–2242 (2016)
30. Van Horn, G., Branson, S., Farrell, R., Haber, S., Barry, J., Ipeirotis, P., Perona, P., Belongie, S.: Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 595–604 (2015)
31. Wang, M., Yang, G.Y., Li, R., Liang, R.Z., Zhang, S.H., Hall, P.M., Hu, S.M.: Example-guided style-consistent image synthesis from semantic labeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1495–1504 (2019)
32. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8798–8807 (2018)
33. Wang, Y., Khan, S., Gonzalez-Garcia, A., Weijer, J.v.d., Khan, F.S.: Semi-supervised learning for few-shot image-to-image translation. In: Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4453–4462 (2020)
34. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3733–3742 (2018)
35. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2223–2232 (2017)