# Cross-Domain Adversarial Autoencoder for Fine Grained Category Preserving Image Translation

Haodi Hou
*State Key Laboratory for Novel Software Technology*
*Nanjing University*
Nanjing, China
hhd@smail.nju.edu.cn

Jing Huo
*State Key Laboratory for Novel Software Technology*
*Nanjing University*
Nanjing, China
huojing@nju.edu.cn

Yang Gao
*State Key Laboratory for Novel Software Technology*
*Nanjing University*
Nanjing, China
gaoy@nju.edu.cn

***Abstract*—Cross-domain image translation attempt to translate images from one domain to another domain, with the content of images preserved. Current approaches treat image's content as the underlying spatial structure, and translation only change image's style of color and texture. These methods can generate realistic results, but may not be able to preserve image's fine grained semantic category information and suffer from the lack of diversity in objects' shapes and viewing angles. In this paper, we propose the problem of fine grained category preserving image translation that aims at preserving image's fine grained category information in cross-domain translation. A novel framework called Cross-Domain Adversarial AutoEncoder (CDAAE) is proposed to solve the problem. CDAAE assumes that cross-domain images have shared content-latent-code space and separate style-latent-code spaces. The content latent code encodes image's basic category information, while the style latent code represents other domain-specific properties, including color, texture, shape, etc. Our experiments evaluate models from aspects of image's quality, diversity as well as category preserving ability, showing CDAAE's advantages over current methods. We also design an algorithm to apply CDAAE to domain adaptation. Experiments on benchmark datasets demonstrate that the proposed method achieves state-of-the-art results.**

***Index Terms*—Autoencoder, Cross-domain image translation, Domain adaptation, Semi-supervised learning**

## I. Introduction

Cross-domain image translation is a significant and challenging task in both multimedia and computer vision. There are many cross-domain image translation applications, such as translation from sketch to photo [1], image to audio [2] and so on. Moreover, cross-domain image translation [3]–[5] can also be applied to domain adaptation. In this paper, we propose to address the problem of fine grained category preserving image translation problem. For this problem, the goal is to generate images of different domains while preserving fine grained category information, such as the identities of faces, the classes of letters, etc. It's much challenging compared with the previous content preserving image translation problem.

There has been many works on cross-domain image transfer. Though they have achieved appealing results, there are several shortcomings. Firstly, they [1], [3], [5]–[8] are unable to extract purely the fine grained category information from images. The work of Huang *et al*. [8] assumes that images can be decomposed into content (domain invariant) and style (domain variant). However, their method can only capture image styles of color and texture, being not able to capture shape or pose related styles. In this way, although they can preserve the content of an image, the content is usually coupled with category, shape and pose information. The work of Liu *et al*. [3] and Zhu *et al*. [7] has the same shortcoming. Unlike their work, we designed a method that is able to extract fine-grained category information from images for cross-domain transfer. As shape and pose are also encoded as style, this leads to more diversity in shape and pose of the generated fine grained images. Besides, most of the current methods lack diversity in generated images. They model image translation as a one-to-one mapping [3]–[5], while image translation is a many-to-many relation in real world. In this paper, we disentangle image's fine grained category information and style into different latent code spaces, so that various images with the same category as input image can be generated by applying different style codes.

In summary, we focus on fine grained category preserving image translation and propose Cross-Domain Adversarial AutoEncoder (CDAAE) to address this problem. The basic assumption of CDAAE is that images from different domains have a shared latent code space for content and separate latent code spaces for styles, where content code decides image's fine grained category and style code influences its color, texture, shape, etc. In order to disentangle category from style, we impose a categorical distribution on content latent code, that can be trained with both labeled and unlabeled data. Since different style codes lead to various appearances of the same category, CDAAE can generate images with more diversity, including color, texture and shape. Besides, the style latent code can also be extracted from sample images, so CDAAE can also be used for sample-guided style transfer. To evaluate

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

image translation methods more comprehensively, we implement experiments measuring image's quality, diversity and whether the category is preserved. Moreover, as our method is able to preserve fine grained category information, we also design a domain adaptation algorithm based on CDAAE, and achieve state-of-the-art accuracy on benchmark datasets. In brief, our contributions are three folds:

- We proposed to use categorical distribution to model the distribution of content code. This leads to a better representation for fine grained category preserving image translation.
- By disentangle the category information from style, the style code of CDAAE can capture not only color and texture, but also shape and pose related style, leading to better diversity in generated images.
- As the proposed method is able to extract category information of images, CDAAE is extended for domain adaptation and achieves state-of-the-art accuracy on benchmark datasets.

## II. Related Works

### A. Image Generation

Variational AutoEncoders (VAEs) [9] and Generative Adversarial Networks (GANs) [10] are the most impressive works on image generation recently. VAEs optimize a lower bound on the log-likelihood of data, and is able to infer the approximate value of the latent code. Instead of KL divergence, Adversarial AutoEncoder (AAE) [11] uses an adversarial training procedure to impose prior distribution on the latent code. Though VAE is comparatively stable to train, the generated image tend to be blurry. GAN is trained through a two-player minimax game, where a generator tries to fool a discriminator while the discriminator tries to distinguish real samples from generated samples. GAN is impressive in generating realistic images, but is unstable and can't do image inference. Effort is made to improve GAN's performance, including stabilizing its training [12], adding reverse network to do image inference [13], [14]. In this paper, we extend autoencoder to deal with cross-domain data, and employ GAN to improve image quality.

### B. Cross-Domain Image Translation

Early works use conditional GANs [1] to translate image into target domain, which needs paired data to train. Triangle GAN [6] combine conditional GAN and Bidirectional GAN by a triangle framework, and achieved semi-supervised image translation. Other works further propose totally unsupervised methods, by making use of semantic features [4], shared weights and latent space [3], [5], as well as cycle consistency [7]. We also focus on image translation without paired data and try to break through some limitations of current methods.

One limitation is that most existing methods treat image translation as a deterministic or unimodal mapping. Consequently, these models can only generate one certain result when input is fixed, while cross-domain images are mostly in a many-to-many relationship. BicycleGAN [15] is able to model continuous and multi-modal distributions, but needs paired data as supervision. In this paper, we attempt to solve the many-to-many problem without paired data. Some concurrent works also try to tackle this problem. Augmented cycleGAN [16] extend CycleGAN [7] with a style code to model diverse styles. While the most similar work to ours is MUNIT [8] that uses affine transformation parameters in normalization layers to represent styles. However, these works can only catch some style of coloring and tend to preserve the exact shape of input image. Our work aims at preserving high-order semantic information and modeling style of shape as well as coloring.

In image translation, it's necessary to keep image's content unchanged. Currently, this is achieved through weights sharing [5], autoencoding with shared latent space [3] or cycle consistency [7], [16]. However, these works tend to preserve all the spatial structure as image's content, which is low-order and hard to evaluate. Their experiments mostly focus on image's quality and lacks evaluation on whether the content is preserved. Our work take image's identity as content so as to preserve high-order semantic information, and our experiments evaluate generated images from three aspect: quality, variety and whether the identity is preserved.

### C. Style Transfer

Style transfer is closely related to cross-domain image translation. Difference is that most style transfer works [17]–[19] transfer style extracted from a style image to the input image, while image translation works transfer style between two sets of images from different domains. As the style code can be extracted from prior distribution as well as a sample image, our model is able to complete both sample-guided style transfer and cross-domain image translation.

## III. Cross-Domain Adversarial Autoencoder

### A. Assumptions

Let $A$ and $B$ be two image domains. $x_A$, $x_B$ are samples from the two domains respectively. Goal of cross-domain image translation is to estimate the two conditionals $P_{A\to B}(x_B|x_A)$ and $P_{B\to A}(x_A|x_B)$ with learned models. When there is no paired data, we can only access samples extracted from the marginal distributions $P_A(x_A)$ and $P_B(x_B)$. As there can be an infinite set of joint distributions yielding the given marginal distributions, we need to make additional assumptions.

Based on the fact that cross-domain images have similar content with different styles, we assume that they have a shared content latent code space $Z^c$ and separate style latent code spaces $Z^s_A$ and $Z^s_B$ respectively. In other words, we postulate there exist functions $\hat{E}^c_A$, $\hat{E}^c_B$, $\hat{E}^s_A$, $\hat{E}^s_B$ and $\hat{G}_A$, $\hat{G}_B$, so that, a pair of corresponding images $(x_A, x_B)$ from the joint distribution $P(x_A, x_B)$ can be encoded as $z^c = \hat{E}^c_A(x_A) = \hat{E}^c_B(x_B)$, $z^s_A = \hat{E}^s_A(x_A)$, $z^s_B = \hat{E}^s_B(x_B)$. In turn, they can be generated by $x_A = \hat{G}_A(z^c, z^s_A)$ and $x_B = \hat{G}_B(z^c, z^s_B)$. This autoencoding procedure is shown in Fig. 1 by solid arrows. We can further impose some prior distributions on $z^s_A$, $z^s_B$

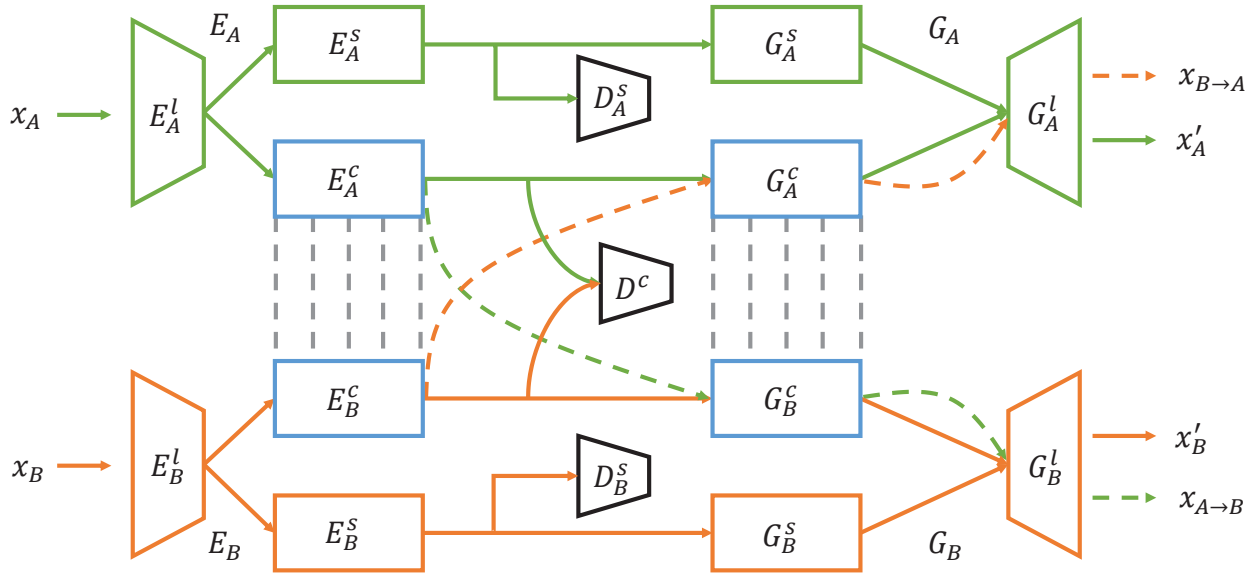


Fig. 1. The proposed CDAAE framework. $E_A$, $E_B$ and $G_A$, $G_B$ are implemented with neural networks. $E_A$ consists of three parts: $E_A^l$ is the lower layers to extract feature map and is shared by $E_A^s$ and $E_A^c$; While $E_A^s$ and $E_A^c$ encode the feature map into style latent code and content latent code respectively. $G_A$ is implemented symmetrically. $G_A^s$ and $G_A^c$ decode the style latent code and content latent code into corresponding feature maps. Then the feature maps are concatenated and fed into $G_A^l$ to produce final images. $E_B$ and $G_B$ are implemented similarly. The shared content latent code space is implemented by sharing weights of $E_A^c$ and $E_B^c$, $G_A^c$ and $G_B^c$ (illustrated by gray dashed lines). $D_A^s$, $D_B^s$ and $D^c$ are adversarial discriminators to tell whether the latent codes are extracted from corresponding prior distributions or not.

and $z^c$, such that cross-domain image translation is done by $x_{A\to B} = \hat{G}_B(\hat{E}^c(x_A), z_B^s)$ and $x_{B\to A} = \hat{G}_A(\hat{E}^c(x_B), z_A^s)$, where $z_A^s$ and $z_B^s$ are extracted from the prior distributions. This is shown in Fig. 1 by dashed arrows. To encode fine grained category information into the content latent code, we assume the content latent code is a one-hot vector with same dimension as category number, and the content part of encoders can be supervisedly trained with categorical labels or unsupervisedly regularized with categorical distribution prior [11]. As for the style latent code, we impose a continuous and simpler Gaussian distribution as prior, since it encodes the comparatively low-dimensional information.

### B. Framework and Training

The proposed framework, as shown in Fig. 1, can be interpreted as a cross-domain version of adversarial autoencoder (AAE). It consists of two domain encoders $E_A$ and $E_B$, two domain generators $G_A$ and $G_B$, and three adversarial discriminators. In particular, $E_A$ is in charge of modeling both $\hat{E}_A^s$ and $\hat{E}_A^c$, while $G_A$ is used to model $\hat{G}_A$ (similar for $E_B$ and $G_B$).

**Weight-sharing.** In order to perform the shared content-latent-code space assumption discussed in Section III-A, we enforce a weight-sharing constraint between the content part of two domains. For encoders, weights of $E_A^c$ and $E_B^c$ are shared; for generators, weights of $G_A^c$ and $G_B^c$ are shared, as illustrated in Fig. 1 by gray dashed lines. These shared weights help to tighten the relation of content code from different domains.

**AAE.** CDAAE's main part is a cross-domain version of AAE [11]. For the reconstruction stream of domain A (shown in Fig. 1 with green solid arrows), encoder $E_A$ firstly maps $x_A$ to codes $z^c$ and $z_A^s$ in latent space $Z^c$ and $Z_A^s$ respectively. Then generator $G_A$ decodes $[z^c, z_A^s]$ to reconstruct the input image. The two discriminators $D^c$ and $D_A^s$ are used to match posterior represented by $E_A$ to the prior distributions through adversarial learning. Therefore, $\{E_A, G_A, D^c, D_A^s\}$ forms an AAE for domain $A$. Since $E_A$ is deterministic, the posterior can be formulated as $q_A^c(z^c|x_A)$ and $q_A^s(z_A^s|x_A)$ and the reconstructed image is $x'_A = G_A(z^c \sim q_A^c(z^c|x_A), z_A^s \sim q_A^s(z_A^s|x_A))$. Similarly for domain $B$, the posterior is $q_B^c(z^c|x_B)$ and $q_B^s(z_B^s|x_B)$ and the reconstructed image is $x'_B = G_A(z^c \sim q_B^c(z^c|x_B), z_B^s \sim q_B^s(z_B^s|x_B))$. Based on the shared content latent code space assumption, $q_A^c(z^c|x_A)$ and $q_B^c(z^c|x_B)$ are matched to a prior $p^c(z^c)$, while $q_A^s(z_A^s|x_A)$ and $q_B^s(z_B^s|x_B)$ are respectively matched to priors $p_A^s(z_A^s)$ and $p_B^s(z_B^s)$. Specifically, we define $p^c(z^c)$ as a categorical distribution [11], $p_A^s(z_A^s)$ and $p_B^s(z_B^s)$ as independent Gaussian distributions. Totally, the cross-domain version of AAE is trained with two losses which are as follows:

$$\begin{aligned} L_{rec} =& E_{x_A\sim P_A(x_A)}[|G_A(E_A(x_A)) - x_A|] \\ &+ E_{x_B\sim P_B(x_B)}[|G_B(E_B(x_B)) - x_B|], \end{aligned} \tag{1}$$

$$
\begin{aligned}
L_{prior} =& E_{x_A \sim P_A(x_A)}[\log(1 - D^c(E_A^c(x_A)))) \\
&+ \log(1 - D_A^s(E_A^s(x_A)))] \\
&+ E_{x_B \sim P_B(x_B)}[\log(1 - D^c(E_B^c(x_B))) \\
&+ \log(1 - D_B^s(E_B^s(x_B)))] \\
&+ E_{z^c \sim p^c(z^c)}[\log(D^c(z^c))] \\
&+ E_{z_B^s \sim p_B^s(z_B^s)}[\log(D_B^s(z_B^s))] \\
&+ E_{z_A^s \sim p_A^s(z_A^s)}[\log(D_A^s(z_A^s))].
\end{aligned} \tag{2}
$$

Eq. (1) is the reconstruction loss. Eq. (2) is the GAN loss to match the posteriors to prior distributions. Here, $E_A^c$ and $E_A^s$ include the part of $E_A^l$ in Fig. 1 (similar for $E_B^c$ and $E_B^s$).

**Content cycle-consistency.** As same content latent code may have different semantic meanings in different domains, we extend the cycle-consistency [7] to content cycle-consistency. In a cycle stream, $x_{A \to B}$ and $x_{B \to A}$ is fed into $E_B$ and $E_A$ to extract their content latent codes. As the content latent code is one-hot vector, we define the content cycle-consistency loss as the cross entropy between the content latent codes of input image and translated image:

$$
\begin{aligned}
L_{cc} =& E_{x_A \sim P_A(x_A), z_B^s \sim p_B^s(z_B^s)}[f_{CE}(E_A^c(x_A), E_B^c(x_{A \to B}))] \\
&+ E_{x_B \sim P_B(x_B), z_A^s \sim p_A^s(z_A^s)}[f_{CE}(E_B^c(x_B), E_A^c(x_{B \to A}))].
\end{aligned} \tag{3}
$$

$f_{CE}$ denotes the function to calculate cross entropy.

**Categorical supervision.** To better disentangle image's fine grained category information from style, we further use categorical supervision loss for data that has a categorical label:

$$
\begin{aligned}
L_{sup} =& E_{x_A \sim P_A(x_A)}[f_{CE}(y_{x_A}, E_A^c(x_A))] \\
&+ E_{x_B \sim P_B(x_B)}[f_{CE}(y_{x_B}, E_B^c(x_B))].
\end{aligned} \tag{4}
$$

Here, $y_{x_A}$ and $y_{x_B}$ are the categorical labels of $x_A$ and $x_B$. To be clear, this loss is used only for data that has a categorical label. Therefore, CDAAE can be trained with both categorically labeled and unlabeled data. We claim that even a few of labeled samples can greatly increase model's category-preserving ability.

In summary, CDAAE is trained by the following optimization:

$$
\begin{aligned}
\min_{E_A, E_B, G_A, G_B} \max_{D^c, D_A^s, D_B^s, D_A, D_B} & \gamma_1 L_{rec} + \gamma_2 L_{cc} \\
& + \gamma_3 Lprior + \gamma_4 L_{sup},
\end{aligned} \tag{5}
$$

where $\gamma_1, \gamma_2, \gamma_3, \gamma_4$ are hyper-parameters to control the weights of these losses.

### C. Domain Adaptation Algorithm

The target of domain adaptation is to generalize the learned model of source domain to a target domain, and label is only available for source domain. Since CDAAE can be trained with both labeled and unlabeled data, it has an innate advantage on domain adaptation. Here, we design an algorithm based on CDAAE to perform domain adaptation.

**Algorithm 1** Domain Adaptation

$S$ is labeled data of source domain, $T$ is unlabeled data of target domain
$t \in [0, 1]$ is the threshold of prediction probability
**for** $i = 1 \cdots pretrain\_steps$ **do**
  $\min L_{sup}$ for $[\boldsymbol{x}_A, y_{x_A}] \in S$
**end for**
**for** $i = 1 \cdots train\_epoches$ **do**
  $T'$={ $[\boldsymbol{x}, y] | \boldsymbol{x} \in T$ and probability of $\boldsymbol{x}$'s label being $y$ $> t$ }
  $\min(L_{prior} + L_{rec} + L_{cc})$ for $[\boldsymbol{x}_A, l_A] \in S, \boldsymbol{x}_B \in T$
  $\min L_{sup}$ for $[\boldsymbol{x}_A, l_A] \in S, [\boldsymbol{x}_B, l_B] \in T'$
**end for**

Thanks to the categorical distribution prior imposed on content latent code (see Section III-B), the content part of encoders in CDAAE can be directly used as a classifier. Though the classifier supervisedly trained for source domain may not have a high accuracy for target domain, its prediction is still meaningful. We can use its prediction as an inexact label for data from target domain. To make best use of supervision from source domain, we further share the weights of $E_A^l$ and $E_B^l$. That is, the content encoder is trained as a classifier for both domain A and B. The algorithm details are shown in Algorithm 1.

## IV. Experiments

### A. Implementation Details and Datasets

As shown in Fig. 1, CDAAE is implemented with neural networks. Encoders and discriminators are mainly comprised of convolution layers and fully connected layers, while generators consist of transposed convolution layers and residual blocks. Dimensions of latent codes are 10 and 8 for content and style respectively. We also use instance normalization to remove style diversity [19]. For more details, we will make our code publicly available in the future.

Our experiments are conducted on digit datasets: the Street View House Number (SVHN) dataset [20], MNIST dataset [21] and USPS dataset [22]. They have 531,131, 60,000, 7291 images for training and 26,032, 10,000, 2007 images for testing respectively. Every image is resized to $32 \times 32$ and grayscale image is replicated three times.

### B. Evaluation Metrics

When evaluate cross-domain translation model, we focus on three aspects of its performance: quality, diversity and whether image's category is preserved. In this paper, we use three metrics to evaluate these abilities respectively.

**Fréchet Inception Distance (FID) [23].** We choose FID to measure image's quality, as it has been shown to be consistent with human evaluation in assessing the realism and variation of the generated images [23]. FID calculates the Wasserstein-2 distance between the generated images and real ones in the feature space of a pretrained Inception-v3 network, so the smaller FID, the better quality.

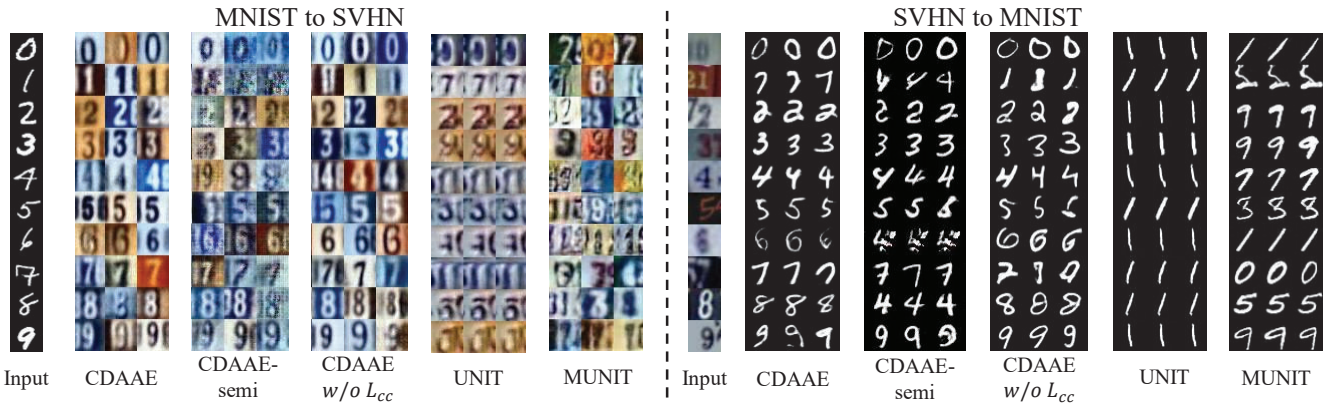


Fig. 2. Cross-domain image translation.

TABLE I
EVALUATION OF CROSS-DOMAIN IMAGE TRANSLATION

| Method | SVHN to MNIST | | | MNIST to SVHN | | |
|---|---|---|---|---|---|---|
| | FID | LPIPS | ACC | FID | LPIPS | ACC |
| CycleGAN | 28.65 | – | 16.53% | 220.7 | – | 6.32% |
| UNIT | 159.4 | 0.0008 | 19.57% | 85.87 | 0.0088 | 31.90% |
| MUNIT | **10.78** | 0.0235 | 18.11% | 46.16 | 0.1974 | 26.29% |
| CDAAE-semi | 28.67 | 0.0778 | 50.92% | 77.96 | **0.1992** | 81.88% |
| CDAAE $w/o$ $L_{cc}$ | 15.66 | **0.0875** | 69.31% | 63.91 | 0.1769 | 96.31% |
| CDAAE | 16.31 | 0.0799 | **88.69%** | **44.63** | 0.1780 | **97.94%** |

TABLE II
EFFECTS OF $\gamma_3$

| $\gamma_3$ | FID | LPIPS | ACC |
|---|---|---|---|
| 0.5 | 18.35 | 0.0799 | 89.10% |
| | 43.01 | 0.1868 | 97.94% |
| 1.0 | 16.31 | 0.0799 | 88.69% |
| | 44.63 | 0.1780 | 97.59% |
| 1.5 | 16.24 | 0.0773 | 88.33% |
| | 43.37 | 0.1764 | 97.60% |

**Learned Perceptual Image Patch Similarity (LPIPS) distance [24].** LPIPS is a weighted $L_2$ distance between deep features of images, and is demonstrated to correlate well with human perceptual similarity [24]. Following Zhu *et al*. [15], we use the average LPIPS distance of image pairs (100 input images and 19 pairs per input) to evaluate diversity. The bigger LPIPS distance, the better diversity.

**Classification Accuracy (ACC).** To evaluate whether image's category is preserved after translation, pretrained classifiers are used to classify the translated images, and accuracy between prediction and ground truth of corresponding input is calculated. Higher accuracy means the category is better preserved. In our experiments, test accuracy of pretrained classifiers for SVHN and MNIST are 93.89% and 99.17% respectively, which is convincing enough to be used as metrics.

### C. Ablation Study

Since there are four components in CDAAE's loss, we firstly conduct experiments to evaluate their effects respectively.

The default hyper-parameters are set as $\gamma_1 = 20$, $\gamma_2 = \gamma_3 = \gamma_4 = 1$. As $L_{rec}$ is a basic component of the total loss, ablation experiments are conducted to evaluate the other three losses. Firstly, to evaluate the effect of labeled data(trained with $L_{sup}$), we only use 1000 categorically labeled images for each class to train in setting of CDAAE-semi. Results are given in Table I It's shown that even with a few of labeled samples, CDAAE can exceed other methods by a large margin in preserving image's category. Secondly, we ablate $L_{cc}$ (denoted as CDAAE $w/o$ $L_{cc}$) to evaluate content cycle consistency. From Table I, it's shown that $L_{cc}$ helps to increase model's ability to preserve image's category. Thirdly, the prior distribution regularization is evaluated by setting $\gamma_3$ to different values. Results are illustrated in Table II. For each row, the upper values are results of SVHN to MNIST, while the lower is for MNIST to SVHN. It is shown that bigger $\gamma_3$ may help to increase image's quality as model is better regularized by prior distributions, but can decrease its diversity and category preserving accuracy.

### D. Results

*1) Cross Domain Image Translation:* Image translation is conducted on SVHN and MNIST. We compare CDAAE with both previous works CycleGAN [7], UNIT [3] and the concurrent work MUNIT [8]. Qualitative results are shown in Fig. 2, and quantitative results are shown in Table 1. As CycleGAN is a deterministic mapping, we only report its FID and ACC.

From both Table 1 and Fig.2, CDAAE is shown to have advantages in category preserving and image diversity. In aspect of **category preserving**, CDAAE can better extract the semantic category information, and make better use of categorical labels, as we impose categorical distribution on the content latent code. Thus CDAAE is the best for ACC. When it comes to **image's diversity** (results of LPIPS), CDAAE is better than others in two aspects. Firstly, the many-to-many design has better diversity than one-to-one mapping. Secondly, CDAAE's style code can encode image's variety both in shape and color, while others can only generate different color

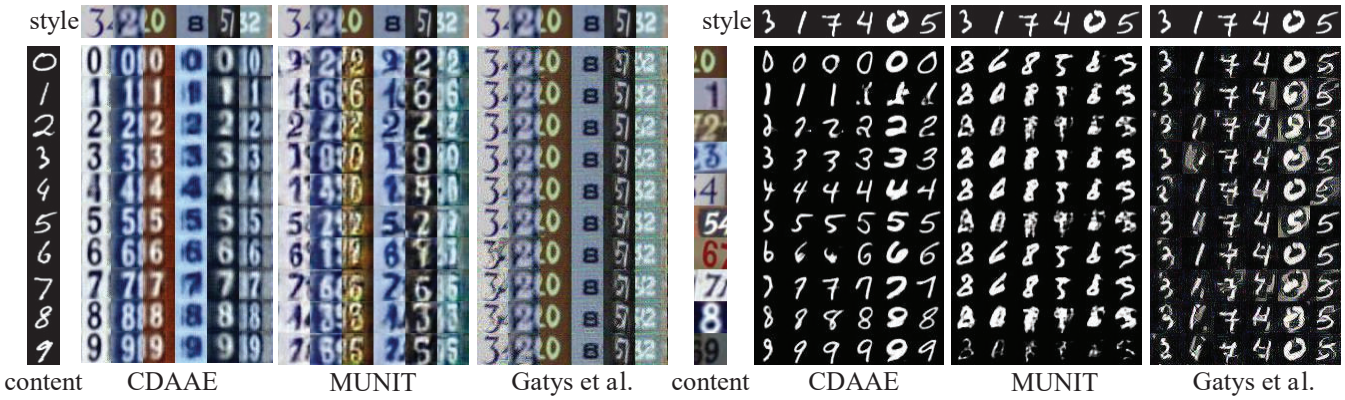


Fig. 3. Sample-guided style transfer.

TABLE III
EVALUATION OF DOMAIN ADAPTATION: ACCURACY ON TEST SPLIT OF TARGET DOMAIN

| Method | SVHN2MNIST | USPS2MNIST | MNIST2USPS |
|---|---|---|---|
| CORAL [25] | – | – | 81.7% [26] |
| MMD [27], [28] | – | – | 81.1% [26] |
| DANN [29] | 73.85% | – | 85.1% [26] |
| DSN [30] | 82.7% | – | 91.3% [26] |
| PixelDA [26] | – | – | 95.9% |
| SA [31] | 59.32% | – | – |
| CoGAN [5] | – | 89.1% | 91.2% |
| DTN [4] | 84.44% | – | – |
| UNIT [3] | 90.53% | 93.58% | **95.97%** |
| CDAAE(ours) | **96.67%** | **96.37%** | 95.91% |

styles. This shown clearly in Fig. 2, especially for SVHN to MNIST, the generated images are more diverse in shapes and writing styles. As for **image quality** (results of FID), CDAAE generates quite good images compared with other methods, and even achieves best quality in translation from SVHN to MNIST.

*2) Sample-Guided Style Transfer:* When style code is extracted from a style image, CDAAE can be used for sample-guided style transfer. This is compared with MUNIT and traditional style transfer method of Gatys *et al*. [17], and results are shown in Fig. 3. As illustrated, CDAAE is better than the other two methods in two aspects. Firstly, CDAAE can preserve the semantic category of content image correctly, since the categorical distribution prior on content code and supervision from categorically labeled data, while MUNIT can't keep the right category. Method of Gatys *et al*. even generates images almost the same as style images, with the content totally wrong. Secondly, CDAAE can catch style including both colors (SVHN style) and shapes (MNIST style), while MUNIT and method of Gatys *et al*. can only catch some color style. This is because CDAAE takes both shape information, color and texture as style, while others refer style only as the rendering on shapes.

*3) Domain Adaptation:* We evaluate the proposed domain adaptation methods (See Section III-C) with the following scenarios: SVHN to MNIST, USPS to MNIST and MNIST to USPS. The threshold of prediction probability is set to $t = 0.85$ (See Algorithm 1). Results are reported in Table III. It's shown that our method achieves comparative accuracy compared with the state-of-the-art work in M2U. In more difficult scenarios: S2M and U2M, we make a considerable progress over the previous state-of-the-art work. As datasets SVHN and MNIST have much more images than USPS, this result illustrates that CDAAE has better ability in making use of unlabeled data. Besides, SVHN and MNIST are much more different than MNIST and USPS, the great progress made by CDAAE in S2M infers that CDAAE generalize better over domain difference.

## V. CONCLUSIONS

We propose the problem of fine grained category preserving image translation and design a novel framework to address it. Experiments show that our method can better preserve image's semantic category and gain much more diversity than existing methods. Besides, domain adaptation algorithm designed based on our framework achieves state-of-the-art result on benchmark datasets. Currently, there is still a limitation on number of image categories in CDAAE, and we plan to address this in the future work.

## ACKNOWLEDGMENT

This work is supported by National Science Foundation of China (61806092), Jiangsu Natural Science Foundation

(BK20180326), and the Collaborative Innovation Center of Novel Software Technology and Industrialization.

## REFERENCES


[1] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5967–5976.

[2] W. Chen, C. Chen, and M. Hu, "Syncgan: Synchronize the latent spaces of cross-modal generative adversarial networks," in *IEEE International Conference on Multimedia and Expo*, 2018, pp. 1–6.

[3] M.-Y. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Advances in Neural Information Processing Systems 30*, 2017, pp. 700–708. [Online]. Available: http://papers.nips.cc/paper/6672-unsupervised-image-to-image-translation-networks.pdf

[4] Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised cross-domain image generation," *arXiv preprint arXiv:1611.02200*, 2016.

[5] M.-Y. Liu and O. Tuzel, "Coupled generative adversarial networks," in *Advances in neural information processing systems*, 2016, pp. 469–477.

[6] Z. Gan, L. Chen, W. Wang, Y. Pu, Y. Zhang, H. Liu, C. Li, and L. Carin, "Triangle generative adversarial networks," in *Advances in Neural Information Processing Systems 30*, 2017, pp. 5247–5256. [Online]. Available: http://papers.nips.cc/paper/7109-triangle-generative-adversarial-networks.pdf

[7] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *The IEEE International Conference on Computer Vision*, 2017, pp. 2223–2232.

[8] X. Huang, M. Liu, S. J. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," *CoRR*, vol. abs/1804.04732, 2018. [Online]. Available: http://arxiv.org/abs/1804.04732

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

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[11] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey, "Adversarial autoencoders," *arXiv preprint arXiv:1511.05644*, 2015.

[12] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017, pp. 2813–2821.

[13] V. Dumoulin, I. Belghazi, B. Poole, A. Lamb, M. Arjovsky, O. Mastropietro, and A. Courville, "Adversarially learned inference," *arXiv preprint arXiv:1606.00704*, 2016.

[14] J. Donahue, P. Krähenbühl, and T. Darrell, "Adversarial feature learning," *arXiv preprint arXiv:1605.09782*, 2016.

[15] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman, "Toward multimodal image-to-image translation," in *Advances in Neural Information Processing Systems 30*, 2017, pp. 465–476. [Online]. Available: http://papers.nips.cc/paper/6650-toward-multimodal-image-to-image-translation.pdf

[16] A. Almahairi, S. Rajeswar, A. Sordoni, P. Bachman, and A. Courville, "Augmented cyclegan: Learning many-to-many mappings from unpaired data," *arXiv preprint arXiv:1802.10151*, 2018.

[17] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2414–2423.

[18] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision*. Springer, 2016, pp. 694–711.

[19] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4105–4113.

[20] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," in *NIPS workshop on deep learning and unsupervised feature learning*, vol. 2011, no. 2, 2011, p. 5.

[21] Y. LeCun, C. Cortes, and C. J. Burges, "Mnist handwritten digit database," *AT&T Labs [Online]. Available: http://yann. lecun. com/exdb/mnist*, vol. 2, 2010.

[22] J. S. Denker, W. Gardner, H. P. Graf, D. Henderson, R. Howard, W. Hubbard, L. D. Jackel, H. S. Baird, and I. Guyon, "Neural network recognizer for hand-written zip code digits," in *Advances in neural information processing systems*, 1989, pp. 323–331.

[23] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Advances in Neural Information Processing Systems 30*, 2017, pp. 6626–6637. [Online]. Available: http://papers.nips.cc/paper/7240-gans-trained-by-a-two-time-scale-update-rule-converge-to-a-local-nash-equilibrium.pdf

[24] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *The IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.

[25] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, ser. AAAI'16. AAAI Press, 2016, pp. 2058–2065. [Online]. Available: http://dl.acm.org/citation.cfm?id=3016100.3016186

[26] K. Bousmalis, N. Silberman, D. Dohan, D. Erhan, and D. Krishnan, "Unsupervised pixel-level domain adaptation with generative adversarial networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 95–104.

[27] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell, "Deep domain confusion: Maximizing for domain invariance," *arXiv preprint arXiv:1412.3474*, 2014.

[28] M. Long, Y. Cao, J. Wang, and M. Jordan, "Learning transferable features with deep adaptation networks," in *Proceedings of the 32nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, F. Bach and D. Blei, Eds., vol. 37. Lille, France: PMLR, 07–09 Jul 2015, pp. 97–105. [Online]. Available: http://proceedings.mlr.press/v37/long15.html

[29] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *Journal of Machine Learning Research*, vol. 17, no. 59, pp. 1–35, 2016. [Online]. Available: http://jmlr.org/papers/v17/15-239.html

[30] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan, "Domain separation networks," in *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds. Curran Associates, Inc., 2016, pp. 343–351. [Online]. Available: http://papers.nips.cc/paper/6254-domain-separation-networks.pdf

[31] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, "Unsupervised visual domain adaptation using subspace alignment," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 2960–2967.