Multimodal and Multiclass Semi-supervised Image-to-Image Translation

Jing Bai[1,2(✉)], Ran Chen[1,2], Hui Ji[1,2], and Saisai Li[1,2]

[1] North Minzu University, Yinchuan 750021, China
baijing@nun.edu.cn
[2] Ningxia Provice Key Laboratory of Intelligent Information and Data Processing, Yinchuan 750021, China

Abstract. In this paper, we propose a multimodal and multiclass semi-supervised image-to-image translation (MM-SSIT) framework to address the dilemma between expensive labeled work and diversity requirement of image translation. A cross-domain adversarial autoencoder is proposed to learn disentangled latent domain-invariant content codes and domain-specific style codes. The style codes are matched with a prior distribution so that we can generate a series of meaningful samples from the prior space. The content codes are embedded into a multiclass joint data distribution by an adversarial learning between a domain classifier and a category classifier so that we can generate multiclass images at one time. Consequently, multimodal and multiclass cross-domain images are generated by joint decoding the latent content codes and sampled style codes. Finally, the networks for MM-SSIT framework are designed and tested. Semi-supervised experiments with comparisons to state-of-art approach show that the proposed framework has the ability to generate high-quality and diversiform images in case of fewer labeled samples. Further experiments in the unsupervised setting demonstrate that MM-SSIT is superior in learning disentangled representation and domain adaption.

Keywords: Image-to-image translation · Semi-supervised · Adversarial auto encoder · Adversarial learning

1 Introduction

Owing to the quick development in AI technology, image-to-image translation has become a compelling topic in recent years [1–3]. Existing approaches usually simplify this problem as a deterministic one-to-one image mapping. However, the cross-domain image translation is multimodal in many scenarios [2, 3]. In this paper, we focus on the multimodal image-to-image translation.

Currently, there are mainly two kinds of image-to-image translation. One of them is supervised [4, 5], which needs paired examples in different domains. Because the requirement is harsh and impractical in many cases, unsupervised image-to-image translation has emerged [2, 3, 6–8]. In order to generate cross-domain images, these methods always assume that the images of two domains share domain-invariant content codes [2], and the content codes share the same data distribution. Unfortunately, this

Y. Zhao et al. (Eds.): ICIG 2019, LNCS 11903, pp. 503–514, 2019.
https://doi.org/10.1007/978-3-030-34113-8_42

assumption is equivalent to the requirement that the cross-domain data must be of single or similar categories. As a result, they fail to generate images between two domains including multiple categories, even the domains containing 0–9.

In this paper, we propose a semi-supervised framework for image-to-image translation, which can achieve multimodal and multiclass image translation with a small number of labeled samples in the absence of paired examples. First of all, we make the same assumptions as MUNIT [2] that the latent space of images can be decomposed into a content space and a style space, and the images in different domains share a common content space but not the style space. Furthermore, we make a different assumption that the images of same categories share the same content distribution. Accordingly, as Fig. 1 shown, we instantiate our semi-supervised translation idea based on the semi-supervised representation learning by introducing the following models:

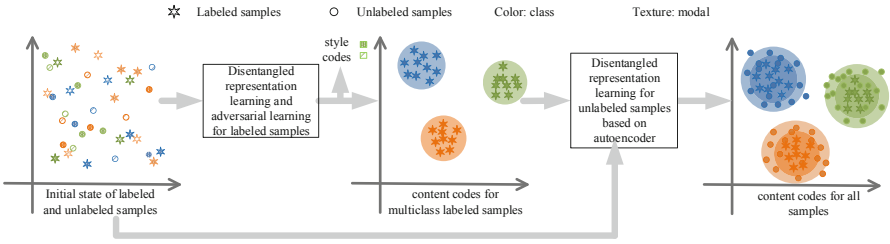


Fig. 1. The multimodal and multiclass semi-supervised representation learning.

Autoencoder (AE). It is used to disentangle the latent codes of style and content, and generate a series of cross-domain images by recombining its content code and the style samples of the target domain.

Adversarial AutoEncoder (AAE). It is added to encoder so as to make the style code of each domain satisfy a particular distribution. Thus, we can generate various style of a domain by sampling from its style distribution.

Adversarial Learning. It is added to encoder so as to embed the domain invariant content attributes into a joint data distribution by an adversarial learning between a domain classifier and a category classifier [9].

This paper makes the following contributions. (1) A Multimodal and Multiclass Semi-supervised Image-to-Image Translation (MM-SSIT) framework is proposed to achieve diversiform image-to-image translation in case of semi-supervised; (2) A novel cross-domain joint data distribution is constructed through the proposed cross-domain adversarial autoencoder, which not only extracts domain invariant content attributes but also captures semantic attributes and makes the content codes of same categories be a cluster; (3) A set of networks for MM-SSIT are designed, which can support semi-supervised image-to-image translation effectively. (4) The experiments on different datasets demonstrate the diversity and superior image quality compared with the state-of-the-art approach.

2 Related Work

2.1 GAN

Generative Adversarial Networks (GANs) have been successfully applied to various computer vision tasks, such as image generation [10–12], image translation [2–8] and semantic segmentation [13]. The work Pix2pix [4] presents conditional adversarial networks as a general solution to image-to-image translation problems, which has achieved remarkable results. However, the work relies on paired examples and only can complete one-to-one mapping. Since then, CycleGAN [6] and DualGAN [7] are proposed to translate an image from a source domain to a target domain in the absence of paired examples by constructing cycle consistency loss. Liu et al. [8] propose an unsupervised image-to-image translation framework based on Coupled GANs [14]. These works have produced good results for image translation, but they can only accomplish one-to-one mapping.

2.2 Multimodal Image Translation

One of these methods can generate a discrete number of outputs by explicitly constructing multimodal codes [15, 16]. The model BicycleGAN [5] can generate continuous and multimodal images. However, the above methods need aligned image pairs for training, which is not available in many tasks. Subsequently, two unsupervised learning works InfoGAN [12] and MUNIT [2] are proposed to generate continuous one-to-more image translation. The only drawback is that these methods require high purity of training data, i.e. implicitly adding a single category restriction to data. This undoubtedly increases its training cost and limits its application scope. Accordingly, Hou et al. propose an image translation framework CDAAE [3], which can generate various samples with a certain input by training in supervised or unsupervised settings. This work provides a useful inspiration for our study.

3 MM-SSIT Framework

3.1 Formulation

Let $x_1 \in \chi_1$ and $x_2 \in \chi_2$ be images from two different domains. Our goal is to learn a more effective joint data distribution from two independent edge data distributions $p(x_1)$ and $p(x_2)$ with fewer labeled samples, then generate multimodal and multiclass cross-domain images for the input image. To solve this problem, we assume that data $x_i \in \chi_i$ can be decoupled independently into a content code $z_i^c \in C_i$ and a style code $z_i^s \in S_i$. Here, for images from a domain i, the content code z_i^c follow the data distribution $q(z_i^c)$, denoting as $z_i^c \sim q(z_i^c)$, and the style code z_i^s follow the data distribution $q(z_i^s)$, denoting as $z_i^s \sim q(z_i^s)$. Then a multiclass joint data distribution of content codes is constructed through an adversarial learning between a domain classifier and a category classifier. Finally, various cross-domain images are generated by joint decoding of the latent content codes and sampled style codes.

3.2 Overall Framework

To address the above issues, Fig. 2 shows the overall framework of MM-SSIT. As shown in the figure, MM-SSIT consists of two parts: an encoder module and a decoder module, which will be described separately below.

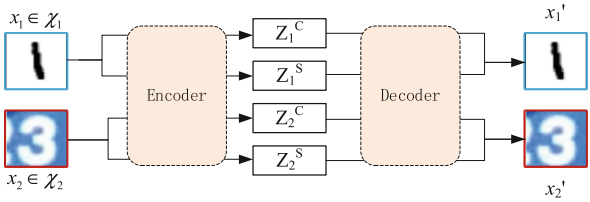


Fig. 2. The overall framework of MM-SSIT.

3.3 Encoder Module

As shown in Fig. 3, inputting an image from a source domain, firstly a disentangled representation learning module is used to decouple its content code and style code, then AAE is used to make the style code satisfy a given distribution, and adversarial learning between a domain classifier and a category classifier is used to make the content features of the same category share the same content distribution.

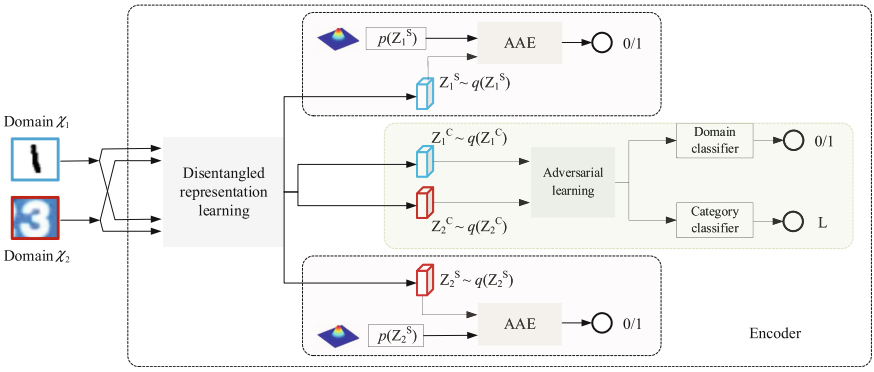


Fig. 3. The encoder framework of MM-SSIT.

Style Encoding Based on AAE. This module is designed to make the style code $z_i^s \sim q(z_i^s)$ satisfy a given distribution $p(z_i^s)$. Therefore, the adversarial loss $L_i^{adv}(i \in \{1,2\})$ between $q(z_i^s)$ and $p(z_i^s)$ can be defined as follows:

$$L_i^{adv} = \log(p(z_i^s)) + \log(1 - q(z_i^s)) \tag{1}$$

Content Encoding Based on Adversarial Learning. With the objective of achieving multiclass cross-domain data generation, this module is designed to construct a multiclass joint data distribution where data from different domains with the same semantic label are in the same cluster. Therefore, the adversarial learning between a domain classifier and a category classifier is introduced into content encoding. Here, the domain classifier is used to determine which domain the received content code is from, thus defining a domain discriminant loss L_{Domain} as formula (2). Furthermore, the category classifier is used to judge category labels of the given images, thus defining a category classifier discriminant loss L_{label} as formula (3). When the training is finished, the joint data distribution can be obtained. In order to ensure a good effect of domain smoothing, the loss functions $L_{cc}^{semi-su}$ and L_{cc}^{un} are defined for the training with fewer labeled data and unlabeled data, respectively.

$$L_{Domain} = \log(q(z_1^c)) + \log(1 - q(z_2^c)) \tag{2}$$

$$L_{label} = F_{CE}(L_1, E_1^c(x_1)) + F_{CE}(L_2, E_2^c(x_2)) \tag{3}$$

$$L_{cc}^{semi-su} = F_{CE}(L_1, E_1^c(x_{1->2})) + F_{CE}(L_2, E_2^c(x_{2->1})) \tag{4}$$

$$L_{cc}^{un} = F_{CE}(E_1^c(x_1), E_1^c(x_{1->2})) + F_{CE}(E_2^c(x_2), E_2^c(x_{2->1})) \tag{5}$$

Where L_i and E_i^c represent the label and the content code of an image $x_i \in \chi_i$, respectively, and F_{CE} (*1,*2) represents the cross-entropy loss between *1 and *2.

3.4 Decoder Module

As shown in Fig. 4, having obtained their content codes and style codes $\left(z_i^c,\ z_i^s\right),\ i \in \{1, 2\}$ for any images $x_i \in \chi_i$, we can reconstruct the original images or generate cross-domain images. In order to achieve generation ability, a pixel-level reconstruction loss between the generated image and the input image is required. Especially, inputting an image $x_i \in \chi_i$, its reconstruction loss $L_i^{rec},\ i \in \{1, 2\}$, can be defined as follows:

$$L_i^{rec} = ||D_i(z_i^c, z_i^s) - x_i||_2 \tag{6}$$

Where, D_i (*1,*2) is the output image by decoding the content code and style code (*1, *2) of the image $x_i \in \chi_i$, $||^*||_2$ represents the L2 regularization norm.

By synthesizing the loss functions of the above stages, the overall loss of the model with fewer labeled samples $L^{semi-su}$ is defined as follows:

$$L^{semi-su} = L_1^{adv} + L_2^{adv} + L_{Domain} + L_{label} + L_{cc}^{semi-su} + L_1^{rec} + L_2^{rec} \tag{7}$$

While the overall loss of the model without any labeled samples L^{un} is defined as follows:

$$L^{un} = L_1^{adv} + L_2^{adv} + L_{Domain} + L_{label} + L_{cc}^{un} + L_1^{rec} + L_2^{rec} \tag{8}$$

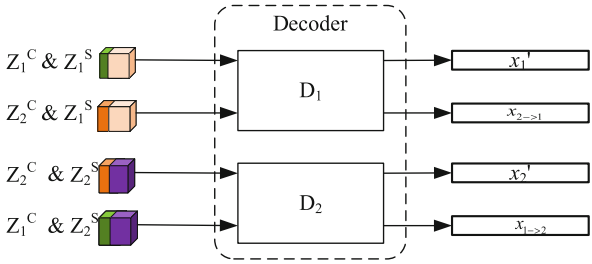


Fig. 4. The decoder framework of MM-SSIT.

4 Network Design

For the proposed framework MM-SSIT, we design a set of networks to achieve semi-supervised multimodal and multiclass image-to-image translation.

4.1 Network for Disentangled Representation Learning Module

Figure 5 shows the designed network for the disentangled representation learning module of MM-SSIT. In this network, the sub-networks for content coding and style coding are composed of 5 layers and 4 layers, while their output are a 8-dimensional style code and a 128-dimensional initial content code, respectively. Because the content code and the style code of an image have the same shallow features, the sub-networks between content coding and style coding share their first two convolution layers. Furthermore, in order to alleviate the gradient disappearance and gradient explosion which aggravate by multilayer neural networks, a batch normalized BN layer (except the last convolution layer of style coding) is added after each convolution layer. More detailed information about every convolution layer is illustrated in Fig. 5.

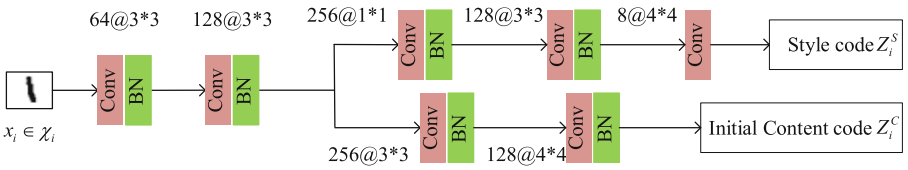


Fig. 5. The network of the disentangled representation learning module.

4.2 Network for Style AAE Module

Figure 6 shows the designed adversarial autoencoder network for the style coding module of MM-SSIT. With a real image code as the negative sample and a random sampling of Normal Distribution as the positive sample, the network is trained to discriminate an input code is true or false through four successive Mlp(256, 64, 16, 1). Finally, the style code disentangled from an image $x_i \in \chi_i$ follows the given Normal Distribution of corresponding domain after training.

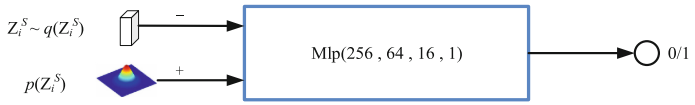


Fig. 6. The network of style AAE module.

4.3 Network for Content Adversarial Learning Module

Figure 7 shows the designed adversarial learning network for the content coding module of MM-SSIT. This network consists of two sub-networks: a category classifier, composing of one Mlp(k) layer and one softmax layer, where k is the number of categories; and a domain classifier, composing of a series of Mlp(256, 128, 64, 64, 2). With initial content codes of two different domains as input, the category classifier is used to classify the images according to their semantic labels, while the domain classifier is designed to discriminate the images' domain, and its output 01 represents domain χ_1 and the output 10 represents domain χ_2. After training, a domain-invariant joint data distribution is constructed, on which the data points from the same classes clustered together while data points from different classes far from each other.

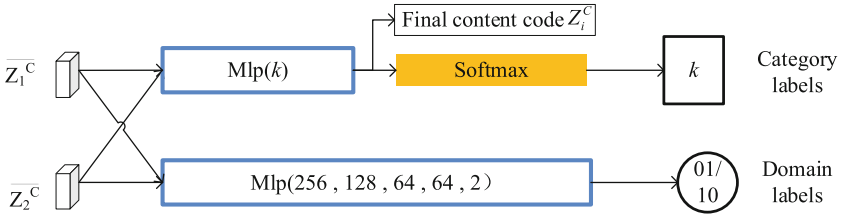


Fig. 7. The network of the content adversarial learning module.

4.4 Network for Decoder Module

As the decoder framework shown in Fig. 8, there are two decoders for MM-SSIT. They share the same network architectures but different training data. The detailed network for each decoder is shown in Fig. 8. Firstly, content codes and style codes are recombined and fed into the decoder. Then one 4 * 4 deconvolution layer with three consecutive 3 * 3 deconvolution layers is recombined to decode images from inputting codes. Finally, the images of data reconstruction or cross-domain generation can be achieved after training is complete. It is noted that a batch normalized BN layer is also added after each deconvolution layer.

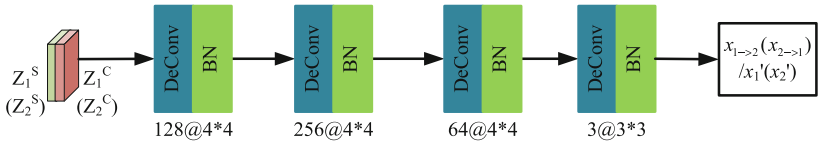


Fig. 8. The network of the decoder module.

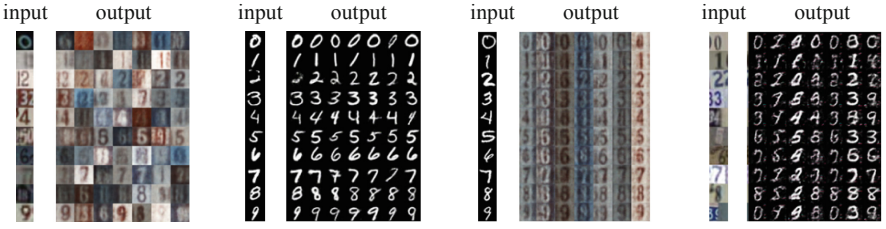
5 Experiments

5.1 Semi-supervised Image-to-Image Translation Experiment

Datasets. We conduct semi-supervised experiments on the dataset MNIST-SVHN. MNIST [17] is composed of 60000 handwritten digit images, while SVHN [18] is composed of 99289 street number image from Street View House Numbers dataset. Both of them are divided into ten categories of 0–9. The digit images from MNIST are gray images of 1 * 28 * 28, while the street number images from SVHN are colored images of 3 * 32 * 32. Therefore, we adjust the digit images to 3 * 32 * 32 three-channel images by data completion and channel expansion before training. In the experiment, 50000 images from MNIST and 73257 images from SVHN are randomly selected for training, and the remaining images are used for testing.

In semi-supervised experiments, there is no need to provide any paired examples, only need to label fewer samples. In this paper, 100 training samples per class are randomly selected for labeling, and the remaining samples are completely unlabeled.

Semi-supervised Experiment on SVHN-MNIST. SVHN (denoting as source domain, s) and MNIST (denoting as target domain, t) are two different domains in this experiment. Figure 9(a)–(d) shows part of the experimental results of s2s, t2t, t2s and s2t, respectively. Here, s2s means the input is an image of source domain SVHN while the outputs are a series of images from the same class but target domain MNIST, the same as t2t, t2s and s2t. The multimodal same-domain translation results in Fig. 9(a) and (b) show that the generated images not only maintain the same content attributes as the input image but also are various and very natural. In addition, the multimodal cross-domain translation results in Fig. 9(c) and (d) are also various, and they show that generated images successfully captures style attributes of the target domain and maintain semantic attributes of the input image. It is noted all the experimental results are achieved in case of nearly 1/62 training data with labels, which demonstrate the effectiveness for semi-supervised image-to-image translation of the proposed framework and designed networks.



(a) s2s (b) t2t (c) t2s (d) s2t

Fig. 9. Semi-supervised image-to-image translation results on SVHN-MNIST.

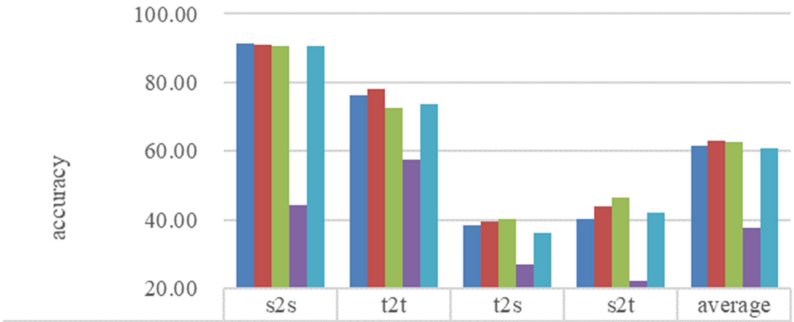
A good image generator should ensure that the generated image and the input image have the same content attributes, i.e. semantic labels. Therefore, the classification accuracy of generated images is calculated and compared with the state-of-art

method CDAAE [3] in the same semi-supervised setting. Table 1 shows all results on MNIST-SVHN, which indicate that the proposed method has an evident advantage over CDAAE for all kinds of image-to-image translations. In addition, whether it is the algorithm in this paper or the CDAAE, the classification accuracies of generated images from the same domain are high, while from different domains are low. This phenomenon illustrates the difficulty of image-to-image translation.

Table 1. Classification accuracy comparisons of four kinds of translations on the MNIST-SVHN. The best performance indicators are marked as bold. (%)

Method	s2s	t2t	t2s	s2t
CDAAE	83.77	72.83	31.06	34.84
MM-SSIT (Ours)	**91.47**	**76.39**	**38.23**	**40.37**

Discussion of Different Partial Weight Sharing Schemes in Decoder. For the two decoders in MM-SSIT, they can share no weights or partial weights. This experiment is designed to evaluate the influence of different weights sharing schemes on the semi-supervised image-to-image translation by comparing classification accuracies. Figure 10 shows the comparison results of our model under the same semi-supervised conditions but different weights sharing schemes between two decoders. In Fig. 10, for s2s and t2t, except for the method with $1^{st}\&2^{nd}\&3^{rd}$ layers weights sharing, all methods achieve similar accuracies; while for s2t and t2s, the method with $1^{st}\&2^{nd}$ layers weights sharing outperforms the other methods. On average, the method only sharing the 1^{st} layer achieves the best performance, which is 2.4%, 0.9%, 67.0% and 4.0% higher than it of the schemes of share-0, share-$1^{st}\&2^{nd}$, share-$1^{st}\&2^{nd}\&3^{rd}$ and share-4^{th}, respectively. The translation results of s2t and t2s for five kinds of weight sharing schemes in Fig. 11 also verify these findings.



	s2s	t2t	t2s	s2t	average
share-0	91.47	76.39	38.23	40.37	61.62
share-1st	91.03	77.94	39.50	43.80	63.07
share-1st&2nd	90.73	72.68	40.08	46.54	62.51
share-1st&2nd&3rd	44.38	57.59	26.91	22.14	37.76
share-4th	90.60	73.57	36.24	42.16	60.64

Fig. 10. Classification accuracy comparisons of four kinds of translations using five kinds of weights sharing schemes of decoders on the MNIST-SVHN.

Discussion of Different Number of Sample Labels. In this section, we will compare the effects of the proposed method using different numbers of labeled samples. Table 2 shows the results of our method by adding labels to 1/62, 1/6, 1/3 and all training data as well as the result of CDAAE [3] with all training data labeled, respectively. Obviously, the classification accuracies increase as the number of labeled samples increases. When labeled samples are 1/62, 1/6 and 1/3, their classification accuracies are 73%, 91% and 96% of the classification accuracy using 100% labeled samples. Furthermore, on average, the performance of our method only using 1/3 labeled samples is comparable to that of CDAAE using 100% labeled samples. These results fully illustrate that the proposed method can achieve high-quality image-to-image translation using fewer labeled samples.

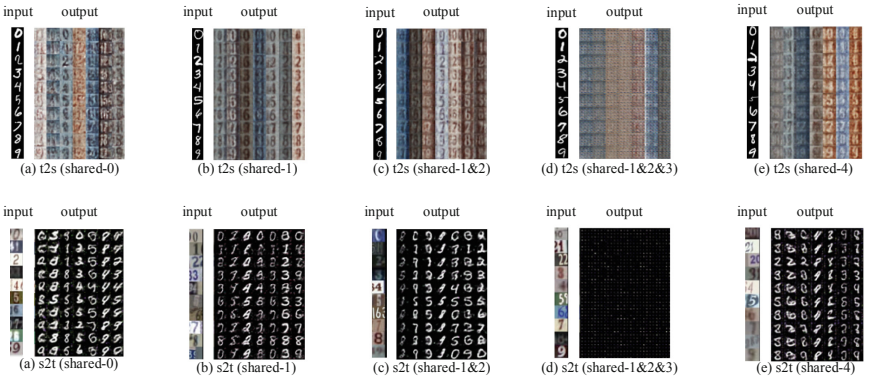


Fig. 11. Semi-supervised image-to-image translation results of s2t and t2s for five kinds of weight sharing schemes on the MNIST-SVHN.

Table 2. Classification accuracy comparisons of different numbers of labeled samples on the MNIST-SVHN. (%)

Method (labeled samples ration in training sets)	s2s	t2t	t2s	s2t	Average
MM-SSIT(1/62)	91.03	77.94	39.5	48.8	64.32
MM-SSIT(1/6)	94.99	88.36	65.91	70.48	79.94
MM-SSIT(1/3)	95.87	90.52	74.38	78.61	84.85
MM-SSIT(1/1)	95.96	90.45	80.78	85.42	85.42
CD-AAE(1/1)	92.03	90.34	78.05	78.81	84.81

5.2 Unsupervised Image-to-Image Translation Experiment

Datasets. We conduct unsupervised experiments on the NIR-VIS and Edges-Shoes. *NIR-VIS* [19]. A face image datasets with two domains including near infrared (NIR) and visible light (VIS) images. This dataset is divided into 724 classes, we select 582 classes with more than five images for both VIS and NIR in our experiment.

In each class, we further select 3 images for training and other 2 images for testing. Here, all images are of 3 * 128 * 128.

Edges-Shoes. A dataset from the pix2pix [4]. For each domain, we randomly select 1000 images for training and 100 images for testing, and resize them into 3 * 128 * 128. In the experiment, we directly use the paired examples provide by pix2pix.

Figure 12(a)–(b) and (c)–(d) show a part of unsupervised image-to-image translation results on VIS-NIR and Edges-Shoes, respectively. In the experiment, the generated image's content code is obtained from the input image, while its style code is sampled from a prior distribution of target domain. It can be seen that for both complex face datasets and shoe datasets with more texture information, the proposed model can achieve good translation results based on the good disentangled representation of contents and styles.

(a) NIR-VIS (b) VIS-NIR (c) Shoes-Edges (d) Edges-Shoes

Fig. 12. Unsupervised image-to-image translation results on VIS-NIR and Edges-Shoes.

6 Conclusion

We present a general framework for semi-supervised image-to-image translation. Our model achieves diversiform and high-quality cross-domain translation results with fewer labeled samples. Future research will further extend the proposed framework to better deal with high-resolution image-to-image translation.

Acknowledgments. This work is supported by National Natural Science Foundation of China (61762003), Natural Science Foundation of Ningxia (2018AAC03124), and Key R&D Program Projects of Ningxia 2019 (Research on Intelligent Assembly Technology Based on Multi-source Information Fusion).

References

1. Zhu, X., Li, Z., et al.: Generative adversarial image super-resolution through deep dense skip connections. In: Computer Graphics Forum (CGF), vol. 37, no. 7, pp. 289–300 (2018)

2. Huang, X., Liu, M.-Y., Belongie, S., Kautz, J.: Multimodal unsupervised image-to-image translation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11207, pp. 179–196. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01219-9_11
3. Hou, H., Huo, J., Gao, Y.: Cross-Domain Adversarial Auto-Encoder (2018). https://arxiv.org/abs/1804.06078. Accessed 17 Apr 2018
4. Isola, P., Zhu, J.Y., Zhou, T., et al.: Image-to-image translation with conditional adversarial networks. In: CVPR 2016, vol. 1, pp. 5967–5976. IEEE Computer Society, Los Alamitos (2017)
5. Zhu, J.Y., Zhang, R., Pathak, D., et al.: Toward multimodal image-to-image translation. In: The 30th Advances in Neural Information Processing Systems, Long Beach, pp. 465–476. Curran Associates (2017)
6. Zhu, J.Y., Park, T., Isola, P., et al.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV 2017, vol. 1, pp. 2242–4421. IEEE Computer Society, Los Alamitos (2017)
7. Yi, Z., Zhang, H., Gong, P.T.M.: DualGAN: unsupervised dual learning for image-to-image translation. In: ICCV 2017, vol. 1, pp. 2868–2876. IEEE Computer Society, Los Alamitos (2017)
8. Liu, M.Y., Breuel, T., Kautz, J.: Unsupervised image-to-image translation networks. In: The 30th Advances in Neural Information Processing Systems, Long Beach, pp. 700–708. Curran Associates (2017)
9. Wang, B., Yang, Y., Xu, X., et al.: Adversarial cross-modal retrieval. In: The 25th ACM International Conference on Multimedia, New York, pp. 157–162 (2017)
10. Goodfellow, I., Pouget-Abadie, J., Mirza, M., et al.: Generative adversarial nets. In: The 27th of Advances in Neural Information Processing Systems, Montreal, pp. 2672–2680. Curran Associates (2014)
11. Zhang, X., Shi, H., Zhu, X., Li, P.: Active semi-supervised learning based on self-expressive correlation with generative adversarial networks. Neurocomputing **345**, 103–113 (2019)
12. Chen, X., Duan, Y., Houthooft, R., et al.: InfoGAN: interpretable representation learning by information maximizing generative adversarial nets. In: The 29th Advances in Neural Information Processing Systems, Barcelona, pp. 2172–2180. Curran Associates (2016)
13. Cai, Q., Xue, Z., Zhang, X., Zhu, X.: A novel framework for semantic segmentation with generative adversarial network. J. Vis. Commun. Image Represent. (JVCI) **58**, 532–543 (2019)
14. Liu, M.Y., Tuzel, O.: Coupled generative adversarial networks. In: The 29th Advances in Neural Information Processing Systems, Barcelona, pp. 469–477. Curran Associates (2016)
15. Chen, Q., Koltun, V.: Photographic image synthesis with cascaded refinement networks. In: ICCV 2017, vol. 1, pp. 1520–1529. IEEE Computer Society, Los Alamitos (2017)
16. Ghosh, A., Kulharia, V., Namboodiri, V., et al.: Multi-agent diverse generative adversarial networks. In: CVPR 2018, vol. 1, pp. 8513–8521. IEEE Computer Society, Los Alamitos (2018)
17. Yann, L., Corinna, C., Christopher, J.B.: MNIST Handwritten Digit Database. AT&T Labs (2010). http://yann.lecun.com/exdb/mnist
18. Yuval, N., Tao, W., Adam, C., et al.: Reading digits in natural images with unsupervised feature learning. In: NIPS Workshop on Deep Learning and Unsupervised Feature Learning, vol. 2011, p. 5 (2011)
19. Li, S., Yi, D., Lei, Z., Liao, S.: The CASIA NIR-VIS 2.0 face database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Los Alamitos, pp. 348–353 (2013)