Independent Encoder for Deep Hierarchical Unsupervised Image-to-Image Translation

Anonymous Author(s)
Affiliation
Address
email

Abstract

The main challenges of image-to-image (I2I) translation are to make the translated image realistic and retain as much information from the source domain as possible. To address this issue, we propose a novel architecture, termed as IEGAN, which removes the encoder of each network and introduces an encoder that is independent of other networks. Compared with previous models, it embodies three advantages of our model: Firstly, it is more directly and comprehensively to grasp image information since the encoder no longer receives loss from generator and discriminator. Secondly, the independent encoder allows each network to focus more on its own goal which makes the translated image more realistic. Thirdly, the reduction in the number of encoders performs more unified image representation. However, when the independent encoder applies two down-sampling blocks, it's hard to extract semantic information. To tackle this problem, we propose deep and shallow information space containing characteristic and semantic information, which can guide the model to translate high-quality images under the task with significant shape or texture change. We compare IEGAN with other previous models, and conduct researches on semantic information consistency and component ablation at the same time. These experiments show the superiority and effectiveness of our architecture. Our code is published on: `https://github.com/Elvinky/IEGAN`.

1 Introduction

I2I translation is an essential topic in the field of computer vision. The goal of I2I translation is to learn mutual mapping functions between two different domains with a bijective relationship[46, 40]. In recent years, there are mainly two approaches of I2I translation. The first approach is based on supervised learning[27, 18, 24, 36], it learns mapping functions from paired image sets. However, in many real-world applications, the workload of collecting paired datasets is extremely heavy, so another approach based on unsupervised learning is proposed[25, 17, 47]. In this approach, due to the lack of mapping relations of paired samples, it's necessary to use additional rules including weight-coupling[25, 26], cycle consistency[46, 21, 40] and identity function[31] to restrict the training of mapping functions.

Most of I2I translation frameworks are composed of generators and discriminators[46, 33, 20]. Like CycleGAN of Figure1, each generator and dicriminator encodes before translating and classifying respectively. However, the encoder of generator has a bottleneck of encoding capability in I2I translation [6, 28]. Compared with the encoder of discriminator directly training through the discrimination loss, gradient received by encoder of generator is back-propagated from the discriminator. This training is indirect for encoder of generator, which causes that the hidden vector learned by an encoder cannot strongly response to the input image. NICE-GAN[6] proposes a solution by removing the encoder of generator, then generator and discriminator share the same encoder of discriminator.

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

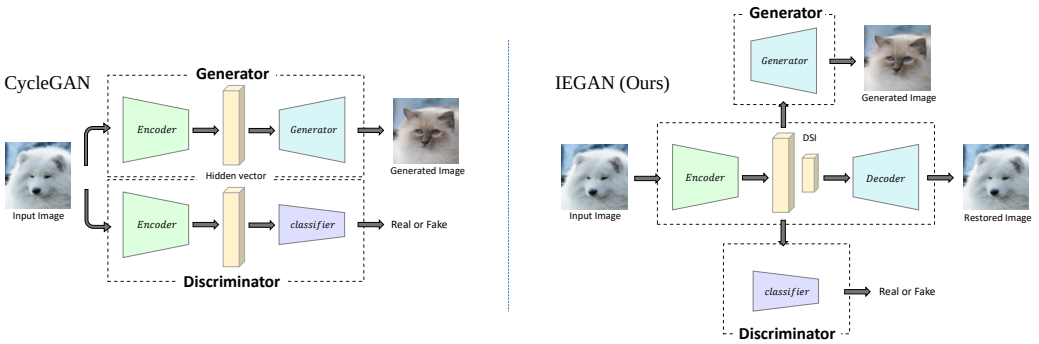

Figure 1: The structural differences between previous GAN models, such as CycleGAN(left), and our IEGAN(right) models. DSI means deep and shallow information space.

Reviewing the goal of each component, the goal of an encoder is to learn the hidden vector that can fully represent the features of input image[2], the decoder can use this hidden vector to restore the input image as much as possible. The ultimate goal of discriminator is to distinguish between the translated image from the source domain and the real image from the target domain[11]. However, discriminators and encoders have different goals, which makes the hidden vector learned by the encoder of discriminator well complete the task of classification but not suitable for task of generation. This is because information that is not conducive to classification won't be learned into the hidden vector.

In view of above mentioned reasons, we propose a novel architecture which refines the goal of each network, as illustrated in Figure1. Specifically, we remove the encoders of generators and discriminators, and introduce an independent encoder, which means that the encoder is no longer affected by other networks. In other words, the generator and discriminator don't need to encode before achieving the goal, and the training of encoder is also independent of the training of other networks. Such kind of independence exhibits three advantages: **I.** Independent encoder can grasp image information more directly and comprehensively. Because this architecture guides the encoder to focus on learning input image representation and ignore the goals of other networks. **II.** The translated image is of higher quality and retains more information from the source domain. **III.** With reduction in the number of encoders in Figure1 decreasing, the number of image representations required for the model is also reduced, which brings the more unified representation.

The performance of previous methods depends on the amount of changes of shape and texture between domains. When independent encoder applies two down-sampling blocks[21, 23], the style transfer can be successfully performed. But it's strenuous to complete tasks with significant shape changes (e.g. The cat is translated into the dog). The hidden vector learned by independent encoder only contains characteristic information (e.g. color and texture) of the input image[38]. To mitigate this problem, we propose deep and shallow information space (DSI) composed of different layers of hidden vectors. In an unsupervised environment, the model obtains the DSI of the input image through an independent encoder, and then asks the decoder to use DSI to restore exactly the same input image. At the same time, DSI merges and superimposes the hidden vectors of different layers and transmits them to the generator and discriminator. In this way, generator and discriminator can use characteristic and semantic information to complete task with significant shape or texture change.

We perform experiments on several popular benchmarks on multiple datasets. Our method outperforms various state-of-the-art counterparts. We further evaluate the independent encoder through semantic information consistency which proves the ability of each model to retain source domain information. In the meantime, ablation studies are conducted to verify the effectiveness of each component.

2 Related work

Generative adversarial networks. GAN[11] has done a large number of practical use cases, such as image generation[42], artwork generation[32], music generation[10], and video generation[35].

In addition, it can also improve image quality[34], image coloring[43], face generation[19], video encoding[37], and other more interesting tasks. GAN has several approaches to improve the authenticity of translated images. The first approach is to improve training stability (e.g. DCGAN[28] used stride convolution and transposed convolution to improve training stability). The second one is large-scale training (e.g. BIGGAN[5] synthesized realistic images by increasing batchsize and truncation techniques). The third one is architectural modifications(e.g. SAGAN[41] added self-attention mechanism to the network). The GAN models mentioned above are all based on probability models, but there are some GAN models based on energy models, such as EBGAN[44] and BEGAN[3]. These models are classified by the reconstruction of input image through the discriminator composed of encoder-decoder.

Unsupervised I2I translation. I2I translation based on unpaired datasets has been widely studied in the field of computer vision since CycleGAN[46] was proposed. U-GAT-IT[20] added AdaILN and CAM[45] to the generator and the discriminator. DeepI2I[38] used a pre-trained discriminator as the encoder of the generator. And NICE-GAN[6] removed the encoder of the generator and used the encoder of the discriminator. These are optimizations for the generator and the discriminator. MUNIT[17] and DRIT[23] decoupled the latent space of images into information of content and style. UNIT[25] and ComboGAN[1] used domain to share latent space. These are researches on image representation. UPD[39] and GANILLA[14] focused on image translation under specific tasks in generating artistic portrait line drawings and illustrations. StarGAN[7] achived multi-domains image translation by adding mask vector to domain labels.

Representation learning. The researches and developments of unsupervised representation learning includes probabilistic models[30], autoencoders[16], and deep networks. The goal of representation learning is to find a better way to express data[2]. A good data representation will greatly improve the efficiency of the model. Representation learning can be implemented in three ways: supervised learning, self-supervised learning and unsupervised learning. Unsupervised learning uses an encoder-decoder network to perform dimensionality reduction and compression on the input data, thereby discarding redundant information, and selecting the most critical concentrated information. In the deep network, the encoder-decoder network constitutes a powerful method of representation learning.

3 Our approach

3.1 Overview

Problem definition. $x \in \{\mathcal{X}\}$ and $y \in \{\mathcal{Y}\}$ are samples respectively in the source domain and target domain. In I2I translation, the ultimate goal is to train the mapping function of $G : x \rightarrow y$ and the inverse mapping function of $F : y \rightarrow x$[46]. In order to train them, some additional constraints are neccessary. For example, in supervised learning, after a paired dataset $\{x_i, y_i\}_{i=1}^{N}$ given, mapping functions are restricted by conditions of $G(x_i) = y_i$ and $F(y_i) = x_i$[18].

But in unsupervised learning, there are only unpaired datasets $\{x_i\}_{i=1}^{N}$ and $\{y_j\}_{j=1}^{M}$. Without restriction of the pairing relation, the function can be mapped to any distributions. To tackle this issue, previous methods put forward additional conditions, such as cycle-consistency $F(G(x_i)) = x_i$ (resp.$G(F(y_j)) = y_j$)[46, 21, 40] and identity-mapping-enforcing $G(y_j) = y_j$ (resp.$F(x_i) = x_i$)[31, 46].

Even if the above conditions guide model to learn mapping functions between two domains, the translated images based on the mapping functions trained by the above conditions are blurred. In other words, the translated image is a mixture of multiple distributions[11]. So GAN introduces two discriminators C_x and C_y, where C_x (resp. C_y) calculates the conditional probability that the sample $F(y_j)$ (resp. $G(x_i)$) matching to the distribution $\mathcal{X}$ (resp. $\mathcal{Y}$).

Independent encoder in I2I translation. In the field of I2I translation, the common GAN models consists of four core networks[6, 20, 39]: two generators G (consists of E_x^G and G) and F (consists of E_y^F and F), two discriminators C_x(consists of E_x^D and C_x) and C_y (consists of E_y^D and C_y). And an encoder is embedded in each core network. We have observed two phenomenas: **I.** Because the input of each network is an image and there are only two image domains, we don't need so many

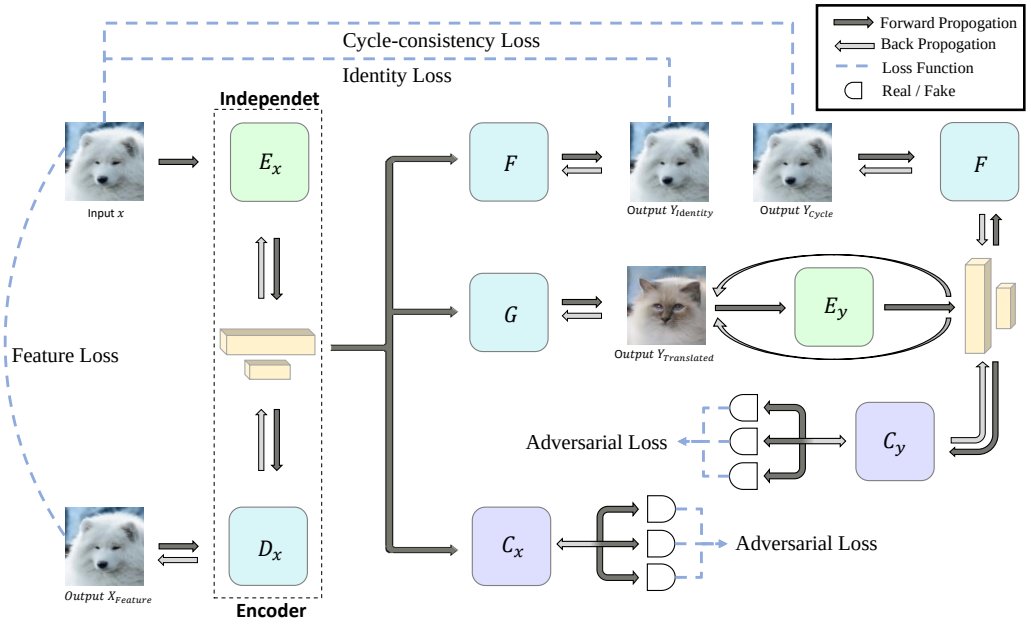

Figure 2: **The structure flowchart of IEGAN.** The structural and purposeful differences between previous GAN models such as CycleGAN(left) and our IEGAN(right) models, DSI means deep and shallow information.

encoders, only two encoders E_x and E_y **II.** The generator needs to encode the image first and then translate it. Also the discriminator encodes first and then classifies the image. However, the generator and discriminator can actually be used without encoding.

As mentioned above, if the encoder is embedded in other networks, the encoding capability of encoder will be limited by the network which it belongs to. This is because training of encoder depends on the back-propagation gradient received by the network. And the gradient depends on the loss function known as the goal of the network.[11] The purpose of encoder is to learn the DSI of the input image, and the goal of discriminator is to map the input image into vector to determine whether the domains are aligned[6]. The vector refers to the output of the generator or the target domain, and the goal of the generator is to translate an image in another domain[46], which means the goal of the encoder is different from the goals of generator and discriminator. If the encoder is in the discriminator, the target of encoder becomes classification, which causes the DSI to lose information irrelevant to classification. If the encoder is in the generator, the loss of encoder is generated by another network. This is an indirect way which causes the DSI not respond strongly to the input image.

So we propose to remove the encoder E_x of each network and establish an encoder-decoder (E_x and D_x) as an independent encoder to be trained, Figure2 illustrates our architecture. The training of the encoder now no longer receives the loss generated by other networks but instead receives the feature loss (Eq. 1) between the output image of the decoder $D_x(E_x(x))$ and the input image x. In this way, the encoder ignores the goal of the generator or discriminator, and focuses on learning the DSI of the input image, thereby ensuring the encoding capacility of encoder. At the same time, the encoder transmits DSI to the generator G and discriminator C_x. When the encoder applies feature loss, DSI also contains more information about the input image. Such an encoder can improve the generating ability of the model, which makes it easier to translate high-quality images.

3.2 Architecture

Independent encoders E_x and E_y. The idea of an independent encoder was inspired by previous work[3, 29, 15, 12]. In these previous works, encoder-decoder network is a powerful method in unsupervised representation learning, the encoder progressively downsamples the input to obtain image information. The decoder uses the learned image information to progressively upsample and restore the input image. When the number of down-sampling blocks is small, the encoder can only

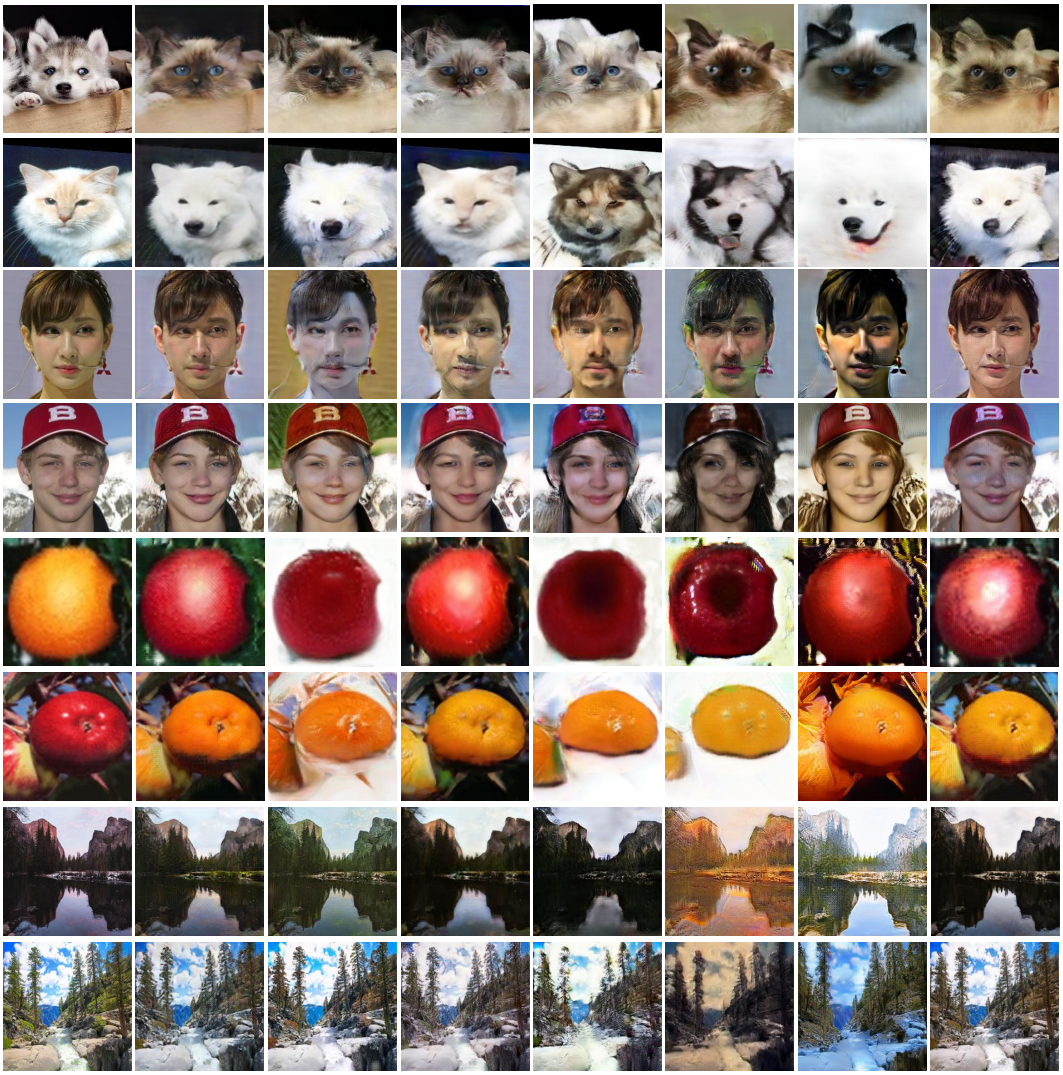

Figure 3: **Examples of translated output on four datasets.** From top to bottom: *cat2dog*, *man2woman*, *apple2orange* and *summer2winter*.

obtain characteristic information. When the number of down-sampling blocks increases, even if encoder obtains semantic information, it is difficult for the decoder to restore a image similar to the input image[38].

By adding the skip connections between encoder and decoder to construct an U-net[29] network, the above problems can be solved ingeniously. In this paper, we apply U-net to build an independent encoder, which means that the independent encoder includes an encoder and a decoder. In order to improve the encoding capability of the encoder, we add linear attention transformer[22] in the encoder.

Most of I2I models only transmit the characteristic information of the input image to generator and discriminator. There are also networks that do other processing, such as decoupling the latent space into content information and style information[17]. As the resolution increases, image information is transmited to the generator through the adapter network layer by layer[38]. We propose an approach that is different from previous models. We extract the information of the input image to DSI, which contains $\mathbb{R}^{H\times W\times C} = \mathbb{R}^{(ch*2^n)\times(ch*2^n)\times(ch*2^{-n})}, n \in \{-3,-2,-1,0\}, ch = 64$. Then the DSI merges and superimposes the hidden vectors of different layers into $\mathbb{R}^{H\times W\times C} =$

Table 1: **Comparison with baselines.** FID and KID×100 for different algorithms. Lower is better.

Method \ Dataset	dog→cat		woman→man		orange→apple		winter→summer	
	FID	KID×100	FID	KID×100	FID	KID×100	FID	KID×100
IEGAN	**43.54**	**0.95**	**138.29**	**3.56**	**132.86**	**4.65**	83.35	1.76
NICE-GAN	48.79	1.58	145.31	4.28	169.79	8.00	**76.44**	**1.22**
U-GAT-IT-light	80.75	3.22	159.73	5.75	134.20	4.92	80.33	1.82
CycleGAN	119.32	4.93	146.97	4.84	142.47	5.46	79.58	1.36
UNIT	59.56	1.94	179.56	9.18	161.25	7.60	95.93	4.63
MUNIT	53.25	1.26	163.39	6.68	186.88	9.09	99.14	4.66
DRIT	94.50	5.20	187.15	10.00	141.07	6.17	78.61	1.69
Method \ Dataset	cat→dog		man→woman		apple→orange		summer→winter	
	FID	KID×100	FID	KID×100	FID	KID×100	FID	KID×100
IEGAN	**38.68**	**0.67**	**136.44**	3.91	**164.58**	**8.70**	97.71	2.07
NICE-GAN	44.67	1.20	139.30	**3.12**	192.19	11.67	**76.03**	**0.67**
U-GAT-IT-light	64.36	2.49	151.06	4.31	179.29	10.17	88.41	1.43
CycleGAN	125.30	6.93	161.62	5.64	181.60	10.42	78.76	0.78
UNIT	63.78	1.94	147.81	4.81	195.24	12.33	112.07	5.36
MUNIT	60.84	2.42	150.94	5.62	213.77	14.04	114.08	5.27
DRIT	79.57	4.57	141.13	3.53	176.13	10.10	81.64	1.27

$\mathbb{R}^{(ch*2^m)\times(ch*2^m)\times(ch*2^{-2m})}, m = 1$, during the up-sampling process, model transmits it to other networks.

Generator and discriminator. Generators G and F are composed of U-GAT-IT-light[20] generators without encoders. The generator contains six layers of residual blocks with AdaILN and two sub-pixel up-sampling layers. The discriminators C_x and C_y are composed of an attention mechanism CAM[45] and a down-sampling network with a multi-scale mechanism[9].

3.3 Loss functions

The training process is consisted of four types of losses: feature loss, adversarial loss, identity reconstruction loss, and cycle-consistency loss. We explain them in detail as follows:

Feature loss. We use average absolute loss to ensure that there is a stable gradient in any input situation to provide a higher quality and more accurate encoding capability. **Except for this loss, the encoders remain unchanged in the following three losses.**

$$\min_{E_x} L_{fea}^{x} = \mathbb{E}_{x\sim\mathcal{X}}[|x - D_x(E_x(x))|_1]. \tag{1}$$

Adversarial loss. The adversarial loss guides the discriminator to distinguish image between source domain and target domain, and make the distribution probability of the output of the generator continuously approach of target domain.

$$\min_{G}\max_{C_y} L_{adv}^{x\to y} = \mathbb{E}_{y\sim\mathcal{Y}}[(C_y(E_y(y)))^2] + \mathbb{E}_{x\sim\mathcal{X}}[(1 - C_y(G(E_x(x))))^2]. \tag{2}$$

Cycle-consistency loss. In order to prevent mode collapse in I2I translation, we apply cycle-consistency loss to have input image translated into target domain, and then translated image back to source domain, which the translated image should be consistent with the input image.

$$\min_{G,F} L_{cyc}^{x\to y} = \mathbb{E}_{x\sim\mathcal{X}}[|x - F(E_y(G(E_x(x))))|_1]. \tag{3}$$

Identity loss. To address the steganography issue[8], we apply identity loss to ensure that G and F no longer have a similar behavior, which is decryption and encryption for hidden vector, so the

Figure 4: **Semantic information consistency.** Compared with other models, IEGAN can better retain the background information of the image under the premise of translating high-quality images.

output distribution of generator can be closer to the distribution of target domain.

$$\min_{F} L_{ide}^{x \to y} = \mathbb{E}_{x \sim \mathcal{X}}[|x - F(E_x(x))|]_1]. \tag{4}$$

Full objective. In summary, final objective of our model is optimized by jointly training independent encoders, generators, and discriminators.

$$\min_{G,F,E_x,Ey} \max_{C_x,C_y} \lambda_1 L_{fea} + \lambda_2 L_{adv} + \lambda_3 L_{cyc} + \lambda_4 L_{ide}, \tag{5}$$

where, $\lambda_1 = 10$, $\lambda_2 = 1$, $\lambda_3 = 10$, $\lambda_4 = 10$. Here, $L_{fea} = L_{fea}^{x} + L_{fea}^{y}$, $L_{adv} = L_{adv}^{x \to y} + L_{adv}^{y \to x}$, $L_{cyc} = L_{cyc}^{x \to y} + L_{cyc}^{y \to x}$ and $L_{ide} = L_{ide}^{x \to y} + L_{ide}^{y \to x}$.

4 Experiments

4.1 Experimental setup

All metrics are obtained through 100K iterations of training on the NVIDIA Tesla P100 GPU. We use the Adam optimizer with 1×10^{-4} weight decay and learning rate, and apply ReLU and leaky-ReLU with a slope of 0.2 as the activation functions of the generator and discriminator respectively. We resize the image to 286×286 and randomly crop to 256×256 for data augmentation.

Baselines. We compare the performance of metrics of our method with six different state-of-the-art I2I translation methods including NICE-GAN[6], U-GAT-IT[20], UNIT[25], MUNIT[17], DRIT[23] and CycleGAN[46], all of which achieves translation between different domains. All models are implemented by the public code on github.

Datasets. We consider four unpaired datasets: *cat2dog*, *man2woman*, *apple2orange*, *summer2winter*. We crop and resize all input images to 256×256 during training, and the output images are also 256×256. The number of splits for the training set and test set of all datasets is based on the following template(train$\mathcal{X}$-$\mathcal{Y}$/test$\mathcal{X}$-$\mathcal{Y}$): *cat2dog* (771-1264/100-100); *man2woman* (1200-1200/115-115); *apple2orange* (995-1019/266-248); *summer2winter* (1231-962/309-238). *cat2dog* dataset is studied in DRIT, *apple2orange* and *summer2winter* dataset are studied in CycleGAN. We created *man2woman* dataset to randomly filter images after classification by gender on FFHQ[19].

Table 2: **Ablation study.** IE: independent encoder; DSI: deep and shallow information space; LAT: add linear attention transformer in IE; × in Components: removing this component; ✓ in Components: containing this component.

Components			dog→cat		cat→dog	
IE	DSI	LAT	FID	KID × 100	FID	KID × 100
×	×	×	54.98	1.85	51.93	1.38
✓	×	×	47.00	1.36	40.81	0.87
✓	✓	×	45.62	**0.92**	39.74	0.79
✓	×	✓	47.80	0.93	40.85	1.06
✓	✓	✓	**43.54**	0.95	**38.68**	**0.67**

Evaluation metrics. For quantitative evaluation, we choose Fréchet Inception Distance[13] (FID) and Kernel Inception Distance[4] (KID) as evaluation metrics. FID calculates the distance between real and translated images in hidden vector given by the features of a convolutional neural network. KID computes the squared Maximum Mean Discrepancy to get the visual similarity of real and translated images. KID has an unbiased estimator, which is more consistent with human evaluation. Low values of FID and KID scores mean the excellent perfomance of GAN.

4.2 Results

Comparisons with state of the arts. We first test IEGAN on four datasets, using six baselines for comparison. We report our key results in Table1 and Figure 3.

Visual analysis. From the perspective of visual analysis, Figure3 shows examples of image translation of different models on different datasets. In general, the images translated by IEGAN are more difficult to distinguish from the images in the target domain, which shows that our model has excellent translation capability at a qualitative level. Both shape and texture are important bases for human perception of images. In terms of shape, the shapes of animals, people, fruits, and landscapes translated by IEGAN are closer to realistic images. For example, facial features of the translated cat in the first row of Figure3 are more realistic than other models. In terms of texture, images translated from models under different architectures will have artifacts to a certain extent. In contrast, our model can better reduce the appearances of artifact.

Metric anaysis. From metric analysis, Table1 shows the FID and KID scores of the above models on the four datasets. In brief, except for summer2winter dataset, the metrics of IEGAN achieve the lowest scores on all datasets and have a significant reduction of FID and KID, especially KID. This shows that our model has an excellent translation capability at a quantitative level. For example, on the popular *cat2dog* dataset, the best KID score of $dog \rightarrow cat$ we obtained was 0.95, which is 0.63 lower than NICE-GAN, and the FID score also dropped from 48.79 to 43.54. Compared to some models that only obtain good scores on the mapping function in one direction, we also reduced the KID from 1.20 to 0.67 on $cat \rightarrow dog$, and the FID from 44.67 to 38.68. We can also notice from the Table1 that CycleGAN can cope with the task of texture change well. UNIT and MUNIT can achieve the goal of shape change. The images translated by DRIT is real but sometimes have nothing to do with input images. The images translated by NICE-GAN and U-GAT-IT have fewer artifacts.

Semantic information consistency. In the process of I2I translation, not every part of image needs to be translated. We refer to the parts that need to be translated and don't need to be translated as subject and background respectively. Even though most of time we only translate the subject, the background is still a part of the image information. The loss of background information from the source domain to the target domain is also a kind of semantic information inconsistency. The semantic information consistency is reflected in the shape, texture, and color of background information. For instance, in the first row of Figure4, a cat is on a table with patterns. After translation, it should be a dog on a table with patterns. In the same example, in the fourth row of Figure4, there is a green label on the apple. After translation, there should be a green label on the orange. Compared with other models, it is obvious that our model has semantic information consistency.

Under tasks of I2I translation, FID and KID are high-quality metrics for evaluating the quality and diversity of translated images, but they cannot fully explain the completeness of translation work. In other words, it is important to translate realistic images, but if semantic information consistency is lost in the translation process, the translation relationship between output and input is weakened[20]. For IEGAN, because the goal of an independent encoder is to focus on learning the DSI of input image, it can preserve as much background information of input image as possible to the generator on the premise of translating high-quality images in order to achieve semantic information consistency.

Ablation study. In the Table2, we next compare the individual impact of each proposed component of the independent encoder in GAN model on the *cat2dog* dataset, and compare the FID and KID scores. We analyze three key components including IE, DSI and LAT. Table2 shows that the application of IE reduces FID and KID by 15% to 37%, which means that an independent encoder can obtain better image representation while also allowing each network to focus on its own goal. This strategy of using independent encoder can improve translation capability of model. In further analysis of DSI, compared with IE, the application of DSI doesn't significantly optimize the quality of the translated image but still lower every metric, which proves the effectiveness of DSI and also means that DSI can make better use of hierarchical information to provide more information to other networks, which makes the quality of translated images better. Finally, we notice that using LAT alone won't produce positive effects, but combined with DSI, it can significantly improve the translation quality of the model. Overall, by merging all components, IEGAN is significantly better than all other variants.

5 Conclusion

In this paper, we propose a novel unsupervised I2I translation architecture, called IEGAN. This architecture allows the encoder to be independent of generator and discriminator. It can more directly and comprehensively grasp the image information. In addition, we introduce a deep and shallow information space based on deep hierarchical I2I translation. The proposed representation allows other networks to obtain the characteristic and semantic information of the input image, which makes the translated image more realistic. Through experiments, we have proved that our model is more superior and effective than previous models.

Broader impact

Our model is mainly used to translate images. The translated images can be applied to business models such as movies, advertisements, games and even virtual reality. Also, this technology can be used to solve the automatic translation of faces or objects. Traditionally, it's a labor-intensive technology, which means that the emergence of this technology has lowered the application threshold for *deepfakes*. Non-professionals can also use this technology to fake information, which may harm the rights of individuals. Seriously, it may endanger the safety of enterprises and the country.

References

[1] Asha Anoosheh, Eirikur Agustsson, Radu Timofte, and Luc Van Gool. Combogan: Unrestrained scalability for image domain translation. In *CVPR Workshops*, pages 783–790, 2018.

[2] Yoshua Bengio, Aaron C. Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828, 2013.

[3] David Berthelot, Tom Schumm, and Luke Metz. BEGAN: boundary equilibrium generative adversarial networks. *CoRR*, abs/1703.10717, 2017.

[4] Mikolaj Binkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD gans. In *ICLR*, 2018.

[5] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *ICLR*, 2019.

[6] Runfa Chen, Wenbing Huang, Binghui Huang, Fuchun Sun, and Bin Fang. Reusing discriminators for encoding: Towards unsupervised image-to-image translation. In *CVPR*, pages 8165–8174, 2020.

[7] Yunjey Choi, Min-Je Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, pages 8789–8797, 2018.

[8] Casey Chu, Andrey Zhmoginov, and Mark Sandler. Cyclegan, a master of steganography. *CoRR*, abs/1712.02950, 2017.

[9] Ishan P. Durugkar, Ian Gemp, and Sridhar Mahadevan. Generative multi-adversarial networks. In *ICLR*, 2017.

[10] Jesse H. Engel, Kumar Krishna Agrawal, Shuo Chen, Ishaan Gulrajani, Chris Donahue, and Adam Roberts. Gansynth: Adversarial neural audio synthesis. In *ICLR*, 2019.

[11] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, pages 2672–2680, 2014.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

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

[14] Samet Hicsonmez, Nermin Samet, Emre Akbas, and Pinar Duygulu. GANILLA: generative adversarial networks for image to illustration translation. *Image Vis. Comput.*, 95:103886, 2020.

[15] Geoffrey E. Hinton, Simon Osindero, and Yee Whye Teh. A fast learning algorithm for deep belief nets. *Neural Comput.*, 18(7):1527–1554, 2006.

[16] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

[17] Xun Huang, Ming-Yu Liu, Serge J. Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, volume 11207, pages 179–196, 2018.

[18] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 5967–5976, 2017.

[19] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019.

[20] Junho Kim, Minjae Kim, Hyeonwoo Kang, and Kwanghee Lee. U-GAT-IT: unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. In *ICLR*, 2020.

[21] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *ICML*, volume 70, pages 1857–1865, 2017.

[22] Nikita Kitaev, Lukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In *ICLR*, 2020.

[23] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *ECCV*, volume 11205, pages 36–52, 2018.

[24] Chunyuan Li, Hao Liu, Changyou Chen, Yunchen Pu, Liqun Chen, Ricardo Henao, and Lawrence Carin. ALICE: towards understanding adversarial learning for joint distribution matching. In *NeurIPS*, pages 5495–5503, 2017.

[25] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *NeurIPS*, pages 700–708, 2017.

[26] Ming-Yu Liu and Oncel Tuzel. Coupled generative adversarial networks. In *NeurIPS*, pages 469–477, 2016.

[27] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014.

[28] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016.

[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, volume 9351, pages 234–241, 2015.

[30] Ruslan Salakhutdinov and Geoffrey E. Hinton. Deep boltzmann machines. In *AISTATS*, volume 5, pages 448–455, 2009.

[31] Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. In *ICLR*, 2017.

[32] Wei Ren Tan, Chee Seng Chan, Hernán E. Aguirre, and Kiyoshi Tanaka. Artgan: Artwork synthesis with conditional categorical gans. In *ICIP*, pages 3760–3764, 2017.

[33] Hao Tang, Dan Xu, Nicu Sebe, and Yan Yan. Attention-guided generative adversarial networks for unsupervised image-to-image translation. In *IJCNN*, pages 1–8, 2019.

[34] Tong Tong, Gen Li, Xiejie Liu, and Qinquan Gao. Image super-resolution using dense skip connections. In *ICCV*, pages 4809–4817, 2017.

[35] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *CVPR*, pages 1526–1535, 2018.

[36] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, pages 8798–8807, 2018.

[37] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. *CoRR*, abs/2011.15126, 2020.

[38] Yaxing Wang, Lu Yu, and Joost van de Weijer. Deepi2i: Enabling deep hierarchical image-to-image translation by transferring from gans. In *NeurIPS*, 2020.

[39] Ran Yi, Yong-Jin Liu, Yu-Kun Lai, and Paul L. Rosin. Unpaired portrait drawing generation via asymmetric cycle mapping. In *CVPR*, pages 8214–8222, 2020.

[40] Zili Yi, Hao (Richard) Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *ICCV*, pages 2868–2876, 2017.

[41] Han Zhang, Ian J. Goodfellow, Dimitris N. Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *ICML*, volume 97, pages 7354–7363, 2019.

[42] Han Zhang, Tao Xu, and Hongsheng Li. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*, pages 5908–5916, 2017.

[43] Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. In *ECCV*, volume 9907, pages 649–666, 2016.

[44] Junbo Jake Zhao, Michaël Mathieu, and Yann LeCun. Energy-based generative adversarial networks. In *ICLR*, 2017.

[45] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, pages 2921–2929, 2016.

[46] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, pages 2242–2251, 2017.

[47] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A. Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *NeurIPS*, pages 465–476, 2017.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] *The scope of this paper is unsupervised I2I translation, and the main contribution is a novel translation architecture, which are shown in the abstract and introduction.*
   (b) Did you describe the limitations of your work? [Yes] *We pointed out in section4.2 of the paper that our model doesn't perform well in the clarity of the translated images on the summer2winter dataset.*
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] *We discussed the impact of our work on the future, see section5 for details.*
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] *We carefully read the ethical review guidelines and ensured that our papers meets these guidelines.*
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] *The assumption of the work of image translation is that there needs to be a bijective relationship between domains, which can be seen in the introduction. For example, the translation from cat to chair and human face to orange is meaningless.*
   (b) Did you include complete proofs of all theoretical results? [Yes] *We compared other models and conducted other experiments to prove our theoretical results.*
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] *The code, data, and instructions to reproduce our experiment are published on the github via URL in abstract.*
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)?[Yes] *We give all the training details of the experiment in section4.1.*
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] *The error of the experiment was mainly caused by the order after the training set was shuffled, but after 100K iterations training, the error will not deviate much. So we did not give the error bars.*
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] *In section4.1, we give the experimental environment and resource consumption.*
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] *We have cited the creators for all data and codes used.*
   (b) Did you mention the license of the assets? [No] *We checked the licenses and cited the datasets. Due to space limitation, those kinds of studies do not state the license information.*
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] *We publish the code on the github via URL in abstract.*
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] *All the data we use are open dataset.*
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] *Face information appeared in the FFHQ dataset, the privacy of which is discussed in the cited article.*
5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]