24th European Conference on Artificial Intelligence - ECAI 2020
Santiago de Compostela, Spain

JND-GAN: Human-Vision-Systems Inspired Generative Adversarial Networks for Image-to-Image Translation

Wenbo Zheng [1] and **Lan Yan** [2] and **Chao Gou** [3,*] and **Fei-Yue Wang** [4]

Abstract. Image-to-image translation aims to learn the mapping between two visual domains. At the beginning of designing the existing image-to-image translation method, it was not considered whether the generated image is realistic or not. In this work, we present a novel approach to address the problem of generating fidelity in the area of image-to-image translation. In particular, humans judge whether an image is realistic or not with unique human vision's feeling rather than paying attention to the real-world semantics. Inspired by this, we propose an effective network loss to capture the pixel-level representations and human vision system information for verisimilar image-to-image translation. To enforce both structural and translation-model consistency during adaptation, we propose a novel Just-Noticeable-Difference loss based on a visual recognition task. The Just-Noticeable-Difference loss not only guides the overall representation to be discriminative, but also enforces our cycle loss before and after mapping between domains. Qualitative results show that our model can generate realistic images on a wide range of tasks without paired training data. For quantitative comparisons, we measure realism with user study and diversity with a perceptual distance metric. We apply the proposed model to domain adaptation and show competitive performance when compared to the state-of-the-art on many datasets.

1 Introduction

Image-to-image translation aims to learn the mapping between two visual domains. The visual domain has presented a more significant challenge from the conversion of non-photo-realistic synthetic data to real images. Although we want to train a large number for models with synthetic data, such as data collected from graphics game engines, these models cannot be generalized to real-world images. The feature level image-to-image methods, such as maximum mean difference, correlation distance, or confrontation discriminator accuracy, solve this problem by aligning features extracted from the network between source (e.g., synthetic) and target (e.g, real) domains without any marked target samples. But these methods cause that a higher level of deep representation alignment may not simulate low-level appearance changes which are critical to the final vision task.

The pixel-level image-to-image model, such as CycleGAN [27], UNIT [17] and DRIT [16], performs a similar distribution alignment, and the source data is converted to the "style" of the target domain, not in the feature space but the original pixel space. This kind of methods produces visually appealing results that preserve local content in natural scenes, but do not take into account the design of final task, which leads that some of the generated images may not look so realistic.

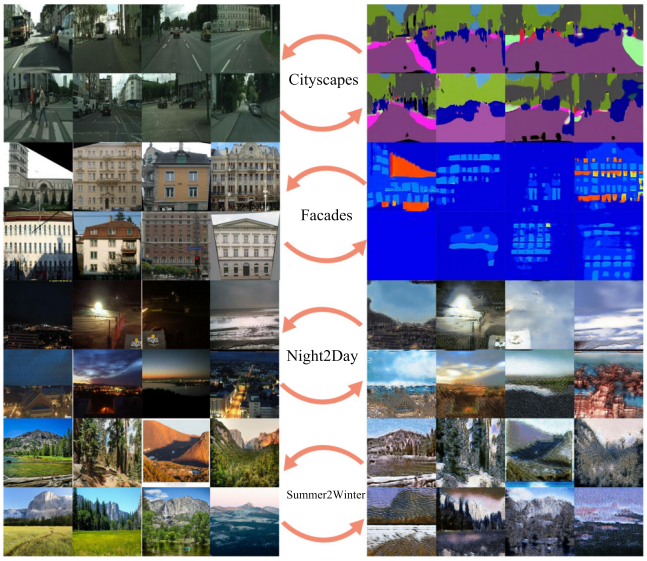


Figure 1. Given any two unordered image collections X and Y, our algorithm learns to automatically "translate" an image from one into the other and vice versa.

In general, the aforementioned two kinds of methods have two disadvantages: one is that aligning the marginal distribution does not enforce the properties of each object. For example, the target feature of a white cloud can be mapped to the source feature of the black cloud. Another is that, in the process of the image to image with GAN technology, the feeling of generative images for human vision is not considered. For example, people feel that generative images are unreal and fake.

To address the above two disadvantages, we consider *why we humans can distinguish an image as real or virtual. Is this related to our vision system?*

The human visual system presents a characteristic that, because of its potential physiological and psychological mechanism, it can only perceive changes in pixels higher than a certain visibility threshold. Just-Noticeable-Difference models refer to the smallest visibility threshold of the human visual system [6]. Now Just-Noticeable-

[1] School of Software Engineering, Xi'an Jiaotong University, Xi'an 710049, China
[2] The State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences
[3] School of Intelligent Systems Engineering, Sun Yat-sen University. *Corresponding author: Chao Gou (e-mail:gouchao@mail.sysu.edu.cn).
[4] The State Key Laboratory for Management and Control of Complex Systems, Institute of Automation, Chinese Academy of Sciences

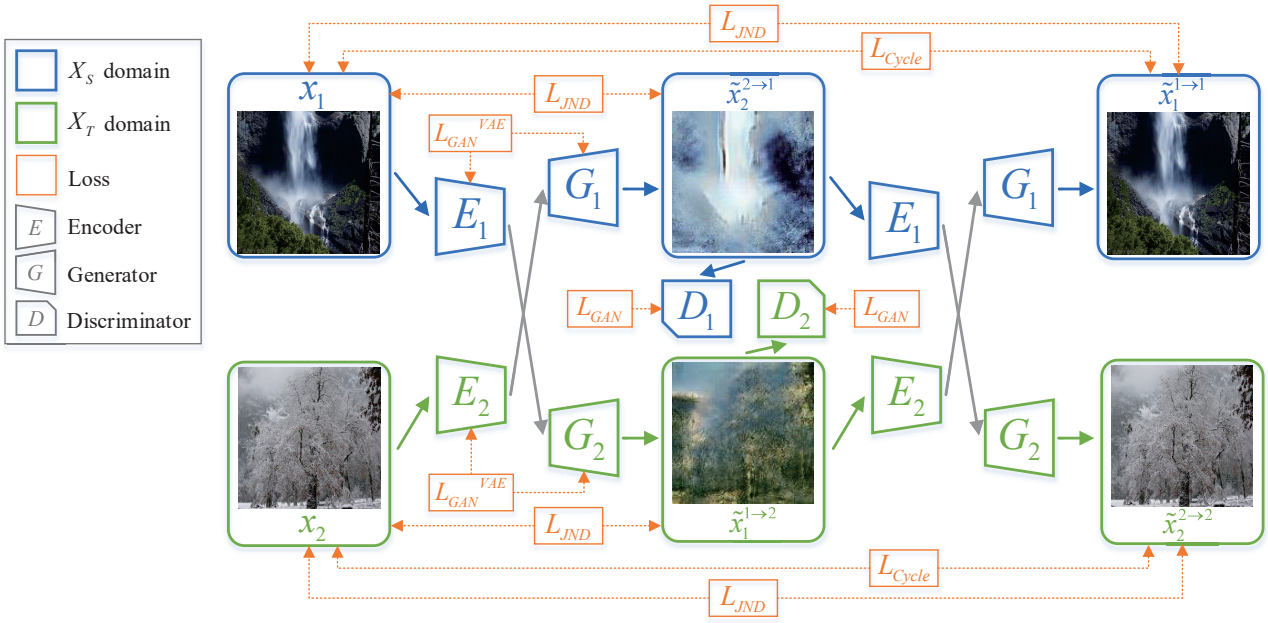


Figure 2. The flow chart of our network during training

Difference models have been widely used in processing perceptual image processing [11]. *Why not using this model in image-to-image technology to make the images more realistic?*

In this paper, we propose a novel image-to-image translation approach inspired by the human visual system called **JND-GAN** to address the problem of the generated images may not look so realistic. We design an effective network loss inspired by the human vision system to combine the pixel-level representations with human vision system information for verisimilar image-to-image translation. We enforce consistency of both structure and model during adaptation using the cycle-consistency loss based on UNIT [17] and Just-Noticeable-Difference loss based on a visual recognition task. The Just-Noticeable-Difference loss both guide the overall representation to be discriminative and enforce Just-Noticeable-Difference consistency before and after mapping between domains. Our approach offers a universal domain adversarial learning model, which combines image-to-image approaches with the strong representational performance of human vision system model. Figure 1 shows our approach of the two testing states. We evaluate our model through extensive qualitative and quantitative evaluation. We measure realism with Fréchet Inception Distance and diversity with the number of statistically-different bins, Jensen-Shannon divergence, and a perceptual distance metric. We also apply our approach to domain adaptation and show competitive performance when compared to others on several datasets.

In short, our main contributions are summarized as follows. We propose just noticeable difference based image-to-image translation approach. Inspired by the human visual systems, we incorporate the Just Noticeable Difference models into GANs to generate more realistic images. Moreover, extensive qualitative and quantitative experiments on benchmark datasets show that our model compares favorably against existing image-to-image models and achieve better results than other state-of-the-art ones.

2 Proposed Method

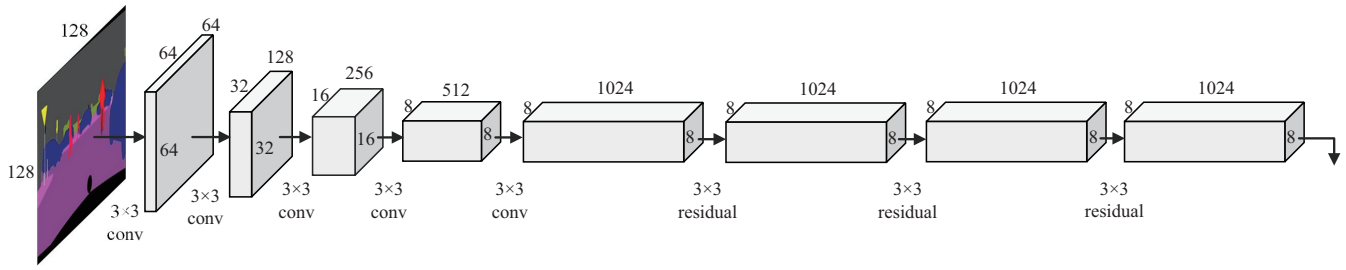
We consider the problem as unsupervised adaptation, where we are provided source data X_S, and target data X_T, but no source and target labels.

We make a shared-latent space assumption. There exists a shared latent code z in a shared-latent space, such that we can recover both images from this code, and we can compute this code from each of the two images.

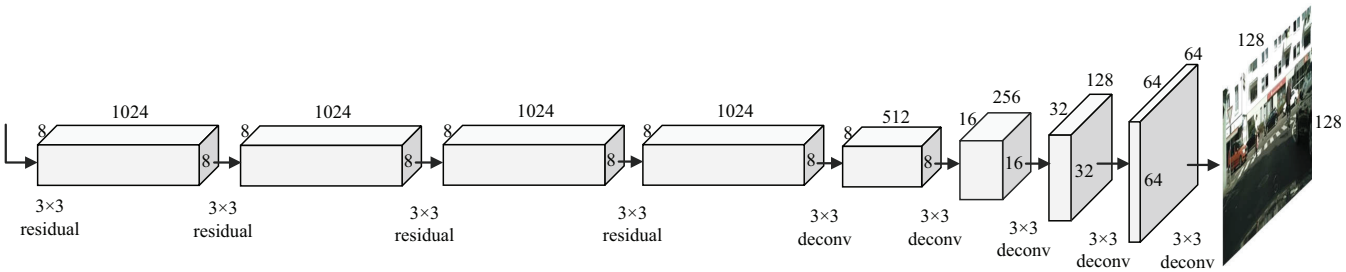
Similar to UNIT [17], our network is built upon variational autoencoders (VAEs) [13] and Cycle generative adversarial network (CycleGAN) [27]. As illustrated in Figure 2, there are two domain image encoders E_1 and E_2, two domain image generators G_1 and G_2, and two domain adversarial discriminators D_1 and D_2 in our network. Take domain $x_1 \in X_S$ and $x_2 \in X_T$ as an example, the encoder E_1 maps images onto a domain-invariant shared space. The generator G_2 generates images conditioned on encoded x_1 image and JND information between x_2 and encoded x_1 image. The discriminator D_2 aims to discriminate between real images and translated images in the domain X_T. Compared to UNIT [17], we use the JND model to increase the self-constraint for domains to make our generated image realistic. Besides,in the following, we first introduce just-noticeable-difference model as one of loss functions of GANs. Then we described our proposed JND-GAN on the basis by the theory of UNIT [17] and introduced the just-noticeable-difference model. Our network architecture is shown in Figure 3.

2.1 Just-Noticeable-Difference Model

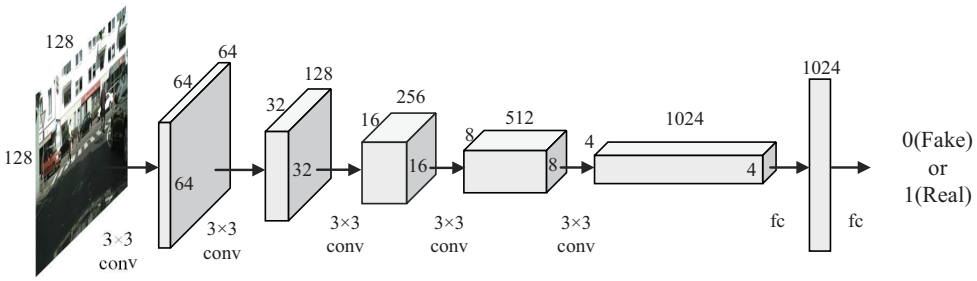
The just noticeable difference (JND) in an image, which reveals the visibility limitation of the human visual system (HVS). The concept of Just-Noticeable-Difference (JND) [24] is widely used in the fields of physiology, psychology of perception, consumer behavior and marketing practice. JND generally refers to a relative threshold in perception by humans. When a change in stimulus value (i.e. change of perceived characteristic) reaches the threshold, the change becomes recognized. When the stimulus value change is below the



(a) The architecture of the encoder



(b) The architecture of the generator



(c) The architecture of the discriminator

Figure 3. The architecture of our generative adversarial networks.

threshold, the change is not being recognized. According to the domain for the visibility threshold to be computed, the existing JND models are usually classified into two kinds, namely, the pixel domain and subband domain JND estimation models. Due to the fast calculation, we choose the pixel domain JND computation. Pixel domain JND computation contains the luminance adaptation modules and contrast masking modules. Given an input image x, the value of the JND model by the work of Wu et al. [24] is as follow:

$$f_{JND}(x) = f_{LA}(x) + f_{CM}(x) - \varpi \times \min\{f_{LA}(x), f_{CM}(x)\} \quad (1)$$

where ϖ is the gain reduction parameter determined by the overlapping between $f_{LA}(x)$ and $f_{CM}(x)$, and here we set $\varpi = 0.3$; $f_{LA}(x)$ is the visibility threshold of the luminance adaptation and $f_{CM}(x)$ is the contrast masking.

2.2 Network Formulation

VAE-GAN Loss. The encoder-generator pair $\{E_1, G_1\}$ constitutes a VAE for the X_S domain, termed VAE_1. Given an input image $x_1 \in X_S$, the VAE_1 first maps x_1 to a code in a latent space z via the encoder E_1 and then decodes a random-perturbed version of the code to reconstruct the input image via the generator G_1. We assume the components in the latent space z are conditionally independent and Gaussian with unit variance. The reconstructed image is $\tilde{x}_1^{1\to1} = G_1(z_1 \sim q_1(z_1|x_1))$. For $\{E_2, G_2\}$ constituting a VAE_2 for X_T, this process is similar as $\{E_1, G_1\}$, $\tilde{x}_2^{2\to2} = G_2(z_2 \sim q_2(z_2|x_2))$. The prior distribution is a zero mean Gaussian $p_\lambda(z) = N(z|0, I)$, where I is an identity matrix and λ is a random vector. Therefore, the VAE-GAN loss function can be written as:

$$L_{GAN}{}^{VAE}(E_1, G_1) = \alpha_1 KL(q(z_1|x_1)||p_\lambda(z)) - \alpha_2 E_{z_1\sim q_1(z_1|x_1)}[\log p_{G_1}(x_1|z_1)] \quad (2)$$

$$L_{GAN}{}^{VAE}(E_2, G_2) = \alpha_1 KL(q(z_2|x_2)||p_\lambda(z)) - \alpha_2 E_{z_2\sim q_2(z_2|x_2)}[\log p_{G_2}(x_2|z_2)] \quad (3)$$

where the hyper-parameters α_1 and α_2 control the weights of the objective terms and the KL divergence terms penalize deviation of the distribution of the latent code from the prior distribution.

Adversarial Loss. For $GAN_1 = \{D_1, G_1\}$, for real images sampled from the first domain, D_1 should output true, while for images generated by G_1, it should output false. G_1 can generate two types of images:

- images from the reconstruction stream $\tilde{x}_1^{1\to1} = G_1(z_1 \sim q_1(z_1|x_1))$;
- images from the translation stream $\tilde{x}_2^{2\to1} = G_1(z_2 \sim q_2(z_2|x_2))$.

For $GAN_2 = \{D_2, G_2\}$, this process is similar as $GAN_1 = \{D_1, G_1\}$. G_2 can generate two types of images:

- images from the reconstruction stream $\tilde{x}_2^{2\to2} = G_2(z_2 \sim q_2(z_2|x_2))$;
- images from the translation stream $\tilde{x}_1^{1\to2} = G_2(z_1 \sim q_1(z_1|x_1))$.

Therefore, the adversarial loss function can be written as:

$$L_{GAN}(E_2, G_1, D_1) = \alpha_3(E_{x_1\sim p_{x_1}}[\log D_1(x_1)] + E_{z_2\sim q_2(z_2|x_2)}[\log(1 - D_1(G_1(z_2))]) \quad (4)$$

$$L_{GAN}(E_1, G_2, D_2) = \alpha_3(E_{x_2\sim p_{x_2}}[\log D_2(x_2)] + E_{z_1\sim q_1(z_1|x_1)}[\log(1 - D_2(G_2(z_1))]) \quad (5)$$

where the hyper-parameter α_3 controls the impact of the adversarial loss function.

Cycle-consistency Constraint Loss. We enforce the Cycle-consistency constraint to further regularize the ill-posed unsupervised image-to-image translation problem, due to the shared-latent space assumption implying the cycle-consistency constraint. We use a VAE-like objective function to model the cycle-consistency constraint, which is given by

$$L_{Cycle_1}(E_1, G_1, E_2, G_2) = \alpha_4(KL(q_1(z_1|x_1)||p_\lambda(z)) + KL(q_2(z_2|{x_1}^{1\to2})||p_\lambda(z)) - \alpha_5 E_{z_2\sim q_2(z_2|{x_1}^{1\to2})}[\log p_{G_1}(x_1|z_2)] \quad (6)$$

$$L_{Cycle_2}(E_2, G_2, E_1, G_1) = \alpha_4(KL(q_2(z_2|x_2)||p_\lambda(z)) + KL(q_1(z_1|{x_2}^{2\to1})||p_\lambda(z)) - \alpha_5 E_{z_1\sim q_1(z_1|{x_2}^{2\to1})}[\log p_{G_2}(x_2|z_1)] \quad (7)$$

where p_{G_1} and p_{G_2} are two Laplacian distributions, the hyper-parameters α_4 and α_5 control the weights of the two different objective terms, and α_4 and α_5 are the same in Eq. (6) and (7).

Just-Noticeable-Difference Loss. A JND model f_{JND} is introduced to the learning image-to-image translation task of GANs. For K-way classification with a cross-entropy loss, this corresponds to

$$\begin{aligned} L_{JND}(f_{JND}, E_1, G_1, E_2, G_2) = \\ -E_{x_1,x_2}\sum_{k=1}^{K}[||{f_{JND}}^{(k)}(G_1(E_2(\tilde{x}_1^{1\to2}))) - {f_{JND}}^{(k)}(x_2)||_2{}^2 + \\ ||{f_{JND}}^{(k)}(G_1(E_1(\tilde{x}_1^{1\to1}))) - {f_{JND}}^{(k)}(x_1)||_2{}^2 + \\ ||{f_{JND}}^{(k)}(G_2(E_1(\tilde{x}_2^{2\to1}))) - {f_{JND}}^{(k)}(x_1)||_2{}^2 + \\ ||{f_{JND}}^{(k)}(G_2(E_2(\tilde{x}_2^{2\to2}))) - {f_{JND}}^{(k)}(x_2)||_2{}^2] \end{aligned} \quad (8)$$

where $||\cdot||$ is the second-ordered norm.

It should be noted that in the original VAE [13], L_{JND} is used to update the encoder. Through the investigation, we found that L_{JND} degrades the quality of the generated images when it is used to update encoder. Hence, we only use L_{JND} when updating our generator.

Full Objective. We jointly solve the learning problems of the VAE_1, VAE_2, GAN_1 and GAN_2 for the image reconstruction streams, the image translation streams, the cycle-reconstruction streams, and JND-target streams:

$$\begin{aligned} \min_{E_1,E_2,G_1}\max_{G_2,D_1,D_2} {L_{GAN}}^{VAE}(E_1, G_1) + {L_{GAN}}^{VAE}(E_2, G_2) \\ + L_{GAN}(E_2, G_1, D_1) + L_{GAN}(E_1, G_2, D_2) \\ + L_{Cycle_1}(E_1, G_1, E_2, G_2) + L_{Cycle_2}(E_2, G_2, E_1, G_1) \\ + L_{JND}(f_{JND}, E_1, G_1, E_2, G_2) \end{aligned} \quad (9)$$

3 Experimental Results

Implementation Details All experiments were conducted using a 4-core PC with an NVIDIA GTX Titan XP GPU, 32GB of RAM, and Ubuntu 16. We use the Adam optimizer [14] with a batch size of 1, for training where the learning rate was set to 0.0001 and momentums were set to 0.5 and 0.999. Each mini-batch consisted of one image from the first domain and one image from the second domain. Our framework has several hyper-parameters. In all experiments, we set the hyper-parameters as follows: $\alpha_1 = \alpha_4 = 0.2$, $\alpha_2 = \alpha_5 = 200$, $\alpha_3 = 20$. Note that we set each of the comparison method and our method to the number of epoch is 3 and the number of iteration is 10000 in order to better compare the performance and results of the algorithm.

Datasets We design Ablation experiments and comparison experiments on several datasets including facades dataset [28], apple2orange dataset [20], cezanne2photo dataset [18], cityscapes dataset [5], edges2handbags dataset [8], edges2shoes dataset [8], horse2zebra dataset [27], maps dataset [23], monet2photo dataset [18], night2day dataset [1], summer2winter dataset [2], ukiyoe2photo dataset [18], and vangogh2photo dataset [18].

Quantitative Evaluation Criteria We evaluated the performance of our algorithm both on in terms of its realism and diversity. Here We use the Learned Perceptual Image Patch Similarity (LPIPS) metric [26] to measure the similarity among images as the quantitative evaluation on the realism of the generated image. The larger the value of LPIPS metric, the more realistic [16, 19] the generated image of this algorithm is. We compute the distance between 1000 pairs of randomly sampled images translated from 100 real images. Further, given a paired data $\{x, y\}$, we can evaluate the diversity of our method's disentanglement on the five paired dataset, by measuring the reconstruction errors of y with $\hat{y} = G_y(E_x(x))$. The reconstruction error is $||y - G_y(E_x(x))||_1$. The smaller the reconstruction error, the better the diversity of this algorithm. Besides, we employ the official implementation of Fréchet Inception Distance (FID) [9], the number of statistically-different bins (NDB) [21] and Jensen-Shannon divergence (JSD) [21]. For NDB and JSD, we use the K-means method on training samples to obtain the clusters. Then the generated samples are assigned to the nearest cluster to compute the bin proportions. As suggested by the author of [21], there are at least 10 training samples for each cluster. Moreover, we follow the work of MSGAN [19] for other experiments' settings.

3.1 Ablation Study

We conduct a series of ablation studies to evaluate the importance of each component in the proposed method. Our ablation experiment is to examine the effectiveness of our JND loss and VAE. Figure 4 and Figure 5 are qualitative results of our ablation experiment. Table 1 and Table 2 are quantitative results of our ablation experiment.

Table 1. The diversity evaluation on our ablation experiment. Using LPIPS metric [26], the larger value, the higher realism.

L_{JND}	VAE	Dataset / Method	facades	apple2orange	cezanne2photo	cityscapes	edges2handbags	edges2shoes	horse2zebra	maps	monet2photo	night2day	summer2winter	ukiyoe2photo	vangogh2photo
×	×	Ours w/o L_JND VAE	0.1498	0.3090	0.1785	0.1645	0.2619	0.3700	0.1359	0.2446	0.2265	0.1737	0.1387	0.1731	0.3721
×	√	Ours w/o L_JND	0.3015	0.3619	0.3389	0.2781	0.2956	0.3842	0.2671	0.2734	0.3406	0.2121	0.1544	0.2079	0.4049
√	×	Ours w/o VAE	0.4080	0.4172	0.4052	0.2854	0.4236	0.3943	0.3061	0.3447	0.3782	0.2708	0.2677	0.3959	0.4386
√	√	Ours	**0.4321**	**0.4324**	**0.4331**	**0.4040**	**0.4349**	**0.4352**	**0.4358**	**0.4365**	0.4369	**0.4372**	**0.4378**	**0.4383**	**0.4393**

Notation "Ours w/o L_{JND}" means our proposed model without the JND loss. "Ours w/o VAE" is our proposed model without the VAE. "Ours w/o L_{JND} and VAE" means our proposed model without the JND loss and VAE. Note that other settings remain unchanged during the Ablation study.

Figure 4. The result of our ablation experiment on edge2handbags dataset

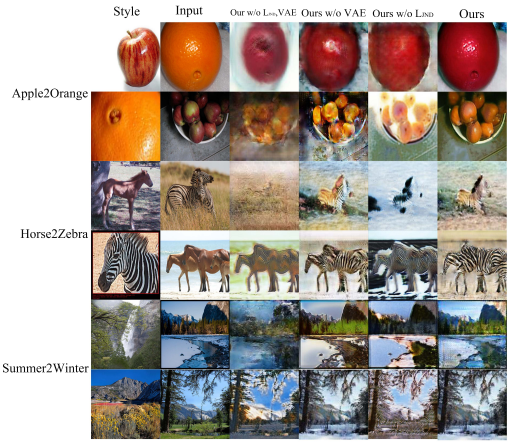


Figure 5. The result of our ablation experiment on unpaired datasets

Qualitative Evaluation on Ablation Study For the "handbag → edge" task shown in Figure 4, it is clear that the result of our method is more realistic than others. When compared "Ours w/o L_{JND} and VAE" with "Ours w/o L_{JND}", we find it that "Ours w/o L_{JND}" can do edge detection, but the results of "Ours w/o L_{JND} and VAE" are unsatisfactory. When compared "Ours w/o VAE" with "Ours", we find it that the results of "Ours" have a richer edge of these handbags than "Ours w/o VAE". These show the VAE is conducive to our network to achieve image-to-image translation. *Therefore, the design of VAE on our JND-GAN is essential.* Compared "Ours w/o L_{JND} and VAE" with "Ours w/o VAE", we find it that the results of "Ours w/o VAE" have richer edges of handbags than "Ours w/o L_{JND} and VAE"; Compared "Ours w/o L_{JND}" with "Ours", it is clear that the results of "Ours" have less noise on the edge of the handbags than "Ours w/o L_{JND}". These show L_{JND} plays a major role in the problem of whether the final generated image is realistic. *Therefore, it is important that we design L_{JND} as one of the loss functions of JND-GAN. By Figure 5, we can get a similar conclusion.*

Table 2. The diversity evaluation on our ablation experiment. Using reconstruction error, the smaller value, the better diversity.

L_{JND}	VAE	Method \ Dataset	facades	cityscapes	edges2handbags	edges2shoes	maps
×	×	Ours w/o L_{JND},VAE	0.17076	0.16335	0.19837	0.14225	0.17347
×	√	Ours w/o L_{JND}	0.22103	0.19167	0.20494	0.15819	0.19685
√	×	Ours w/o VAE	0.23305	0.19205	0.20896	0.18439	0.20273
√	√	Ours	**0.13481**	**0.13487**	**0.13493**	**0.13495**	**0.13498**

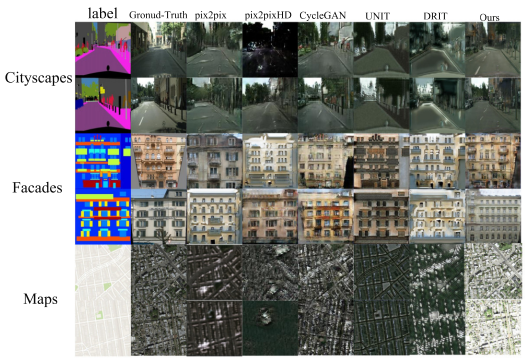


Figure 6. The comparison results on "label to image" task.

Quantitative Evaluation on Ablation Study To further clarify the contribution of the design of VAE and L_{JND}, we analyze the following two aspects: comparison of tasks whether using VAE and comparison of tasks whether using L_{JND}:

Comparison of Tasks Whether Using VAE From Table 1, "Ours w/o L_{JND}" is 0.1136 higher than "Ours w/o L_{JND} and VAE". Also, "Ours" is 0.1486 higher than "Ours w/o VAE". From two these points, we find it that "VAE" is helpful to our network to achieve image-to-image translation. *Therefore, it is necessary for the design of VAE on our JND-GAN. By Table 2, we can get a similar conclusion.*

Comparison of Tasks Whether Using L_{JND} From Table 1, "Ours w/o VAE" is 0.1209 higher than "Ours w/o L_{JND} and VAE". Also, "Ours" is 0.1599 higher than "Ours w/o L_{JND}". From two these points, we find it that L_{JND} plays a major role to influence our network to get the final realistic generated image. *Therefore, it is crucial that we design L_{JND} as one of the loss functions of JND-GAN. By Table 2, we can get a similar conclusion.*

3.2 Comparison with State-of-the-Art Methods

We design comparison experiments and also perform domain adaptation on the classfication task with MNIST [15] to MNIST-M [7]. We perform the evaluation on the following algorithms: Pix2Pix [10], Pix2PixHD [22], CycleGAN [27], DualGAN [25], DiscoGAN [12], UNIT [17] and DRIT [16].

3.2.1 *Qualitative Evaluation*

Comparison Results on Paired Datasets As Pix2Pix [10], there are two tasks on paired datasets, one is to transform the label to image called "label to image", and the other is to transform the image to the label called "image to label". For the "label to image" task, Figure 6 shows the results of Pix2Pix [10], Pix2PixHD [22], CycleGAN [27], UNIT [17], DRIT [16] and ours. On the Cityscapes dataset, the ground in the resulting image of Pix2Pix and CycleGAN is bulging, which is inconsistent with the ground-truth. The resulting image of UNIT and DRIT is fuzzy. But the results of our method are very close to the ground-truth. For the "image to label" task, Figure 7 shows the results of CycleGAN [27], UNIT [17], DRIT [16] and ours. On the

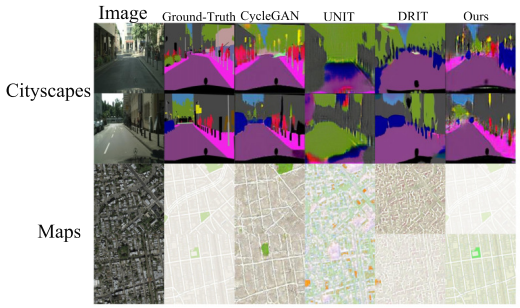


Figure 7. The comparison results on "image to label" task.

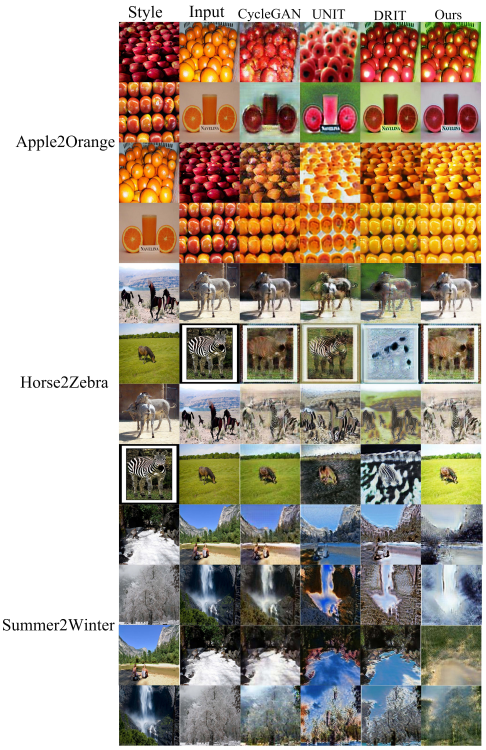


Figure 8. The comparison results on unpaired datasets.

Cityscapes dataset, the results of our method and CycleGAN [27] are significantly better than UNIT [17] and DRIT [16]. The results of our method and CycleGAN [27] are very close to the image of Ground-Truth. In general, these results means our method is better than others, and the result of our method is realistic.

Comparison Results on Unpaired Datasets As CycleGAN [27] and UNIT [17], the task is to transform the style to the image on unpaired datasets. On apple2orange dataset, as we have seen, in the images of the third and fourth lines, our method is superior to other methods. Moreover, on horse2zebra dataset, as we can see, in the image of the eighth line, our method preserves the effect of the input image grassland, so it is more realistic than the results of other methods. Furthermore, on summer2winter dataset, as is shown in the image of the eleventh line, our method is better in term of the forest road, and the other methods are wrong as the sky, so it is more realistic than the results of other methods. In general, these results mean our method is better than others, and the result of our method is realistic.

Table 3. Realism evaluation comparison. Using FID [9], the smaller value, the higher realism.

Dataset \ Method	Ours	DRIT [16]	UNIT [17]	CycleGAN [27]	DiscoGAN [12]	DualGAN [25]	Pix2PixHD [22]	Pix2Pix [10]
facades	**94.13**	100.17	101.23	108.70	111.23	121.21	139.19	139.19
apple2orange	**44.04**	50.23	51.19	51.39	51.72	53.96	-	-
cezanne2photo	**45.29**	51.67	52.15	53.38	53.68	54.37	-	-
cityscapes	**91.53**	98.45	99.08	99.56	103.62	104.97	107.21	107.28
edges2handbags	**126.13**	132.34	133.85	133.93	134.00	134.52	136.33	140.51
edges2shoes	**16.90**	23.74	24.58	26.93	29.24	29.39	29.93	30.09
horse2zebra	**49.96**	56.46	57.13	58.23	58.76	58.79	-	-
maps	**94.14**	100.23	101.24	101.69	101.80	103.25	103.58	103.81
monet2photo	**28.00**	34.67	35.53	35.99	36.30	36.94	-	-
night2day	**29.43**	35.56	36.06	36.10	36.28	38.17	-	-
summer2winter	**47.73**	54.24	55.25	55.30	56.44	56.45	-	-
ukiyoe2photo	**41.03**	47.89	48.53	49.01	49.57	49.60	-	-
vangogh2photo	**52.13**	58.67	59.02	59.31	59.39	59.54	-	-

Table 4. Diversity evaluation comparison. Using NDB [21], the smaller value, the higher realism.

Dataset \ Method	Ours	DRIT [16]	UNIT [17]	CycleGAN [27]	DiscoGAN [12]	DualGAN [25]	Pix2PixHD [22]	Pix2Pix [10]
facades	**10.93**	13.10	14.82	15.01	20.73	22.24	22.93	31.24
apple2orange	**39.92**	42.00	42.98	43.66	44.77	46.85	-	-
cezanne2photo	**38.62**	41.20	42.53	46.88	51.30	52.87	-	-
cityscapes	**10.72**	13.50	14.62	14.65	18.08	20.27	107.21	107.28
edges2handbags	**15.03**	17.56	18.56	19.48	20.12	29.41	22.37	27.12
edges2shoes	**14.95**	17.23	18.73	25.45	29.30	42.86	43.62	44.16
horse2zebra	**42.34**	44.78	45.46	45.90	46.36	50.00	-	-
maps	**18.56**	20.78	21.29	22.70	26.58	26.90	27.16	36.30
monet2photo	**40.98**	43.78	44.76	45.81	48.99	49.36	-	-
night2day	**40.35**	42.89	43.21	44.01	46.80	49.26	-	-
summer2winter	**23.46**	25.60	26.15	26.51	35.25	35.62	-	-
ukiyoe2photo	**21.27**	23.87	24.49	26.42	30.31	36.69	-	-
vangogh2photo	**24.45**	26.76	27.43	28.06	41.03	43.23	-	-

Table 5. Diversity evaluation comparison. Using JSD [21], the smaller value, the higher realism.

Dataset \ Method	Ours	DRIT [16]	UNIT [17]	CycleGAN [27]	DiscoGAN [12]	DualGAN [25]	Pix2PixHD [22]	Pix2Pix [10]
facades	**0.0420**	0.0710	0.0710	0.0710	0.0980	0.1221	0.0740	0.7510
apple2orange	**0.0944**	0.1236	0.1241	0.1941	0.2401	0.2621	-	-
cezanne2photo	**0.0979**	0.1270	0.1271	0.1850	0.1910	0.2231	-	-
cityscapes	**0.0541**	0.0833	0.0841	0.1131	0.1221	0.1871	0.1960	0.2130
edges2handbags	**0.0967**	0.1267	0.1270	0.1790	0.2210	0.2221	0.2450	0.2510
edges2shoes	**0.2295**	0.2587	0.2590	0.2611	0.2621	0.2760	0.2801	0.2931
horse2zebra	**0.0385**	0.0678	0.0680	0.1280	0.1581	0.2301	-	-
maps	**0.0852**	0.1145	0.1151	0.1191	0.1251	0.1440	0.1621	0.1731
monet2photo	**0.0387**	0.0679	0.0680	0.0880	0.1701	0.2470	-	-
night2day	**0.0873**	0.1166	0.1171	0.1460	0.2450	0.2600	-	-
summer2winter	**0.0363**	0.0660	0.0661	0.0950	0.1491	0.1961	-	-
ukiyoe2photo	**0.1157**	0.1456	0.1461	0.1650	0.2100	0.2371	-	-
vangogh2photo	**0.1395**	0.1693	0.1700	0.1951	0.2140	0.2281	-	-

Table 6. The realism evaluation comparison. Using LPIPS metric [26], the larger value, the higher realism.

Dataset \ Method	Ours	DRIT [16]	UNIT [17]	CycleGAN [27]	DiscoGAN [12]	DualGAN [25]	Pix2PixHD [22]	Pix2Pix [10]
facades	**0.4321**	0.4276	0.4080	0.4134	0.2307	0.2213	0.4116	0.4068
apple2orange	**0.4324**	0.4234	0.4153	0.4155	0.2309	0.2236	-	-
cezanne2photo	**0.4331**	0.4241	0.4169	0.4218	0.2348	0.2281	-	-
cityscapes	**0.4340**	0.4244	0.4125	0.4246	0.2418	0.2343	0.4127	0.4191
edges2handbags	**0.4349**	0.4245	0.4129	0.4234	0.2498	0.2419	0.4116	0.4021
edges2shoes	**0.4352**	0.4254	0.4130	0.4138	0.2531	0.2517	0.4128	0.4029
horse2zebra	**0.4358**	0.4260	0.4133	0.4247	0.2545	0.2615	-	-
maps	**0.4365**	0.4267	0.4137	0.4126	0.2586	0.2633	0.4043	0.4039
monet2photo	**0.4369**	0.4272	0.4141	0.4264	0.2626	0.2713	-	-
night2day	**0.4372**	0.4276	0.4145	0.4127	0.2662	0.2809	-	-
summer2winter	**0.4378**	0.4283	0.4149	0.4273	0.2742	0.2887	-	-
ukiyoe2photo	**0.4383**	0.4288	0.4157	0.4276	0.2745	0.2925	-	-
vangogh2photo	**0.4393**	0.4293	0.4165	0.4184	0.2781	0.2998	-	-

Table 7. The diversity evaluation comparison. Using reconstruction error, the smaller value, the better diversity.

Test Dataset \ Method	Ours	DRIT [16]	UNIT [17]	CycleGAN [27]	DiscoGAN [12]	DualGAN [25]	Pix2PixHD [22]	Pix2Pix [10]
facades	**0.13481**	0.20672	0.21765	0.17791	0.39986	0.38872	0.15655	0.172508
cityscapes	**0.13487**	0.20674	0.21775	0.17797	0.39988	0.38877	0.15664	0.172524
edges2handbags	**0.13493**	0.20676	0.21780	0.17802	0.39995	0.38880	0.15667	0.172586
edges2shoes	**0.13495**	0.20682	0.21785	0.17803	0.40001	0.38885	0.15672	0.172646
maps	**0.13498**	0.20683	0.21788	0.17812	0.40002	0.38892	0.15672	0.172672

Table 8. Domain adaptation results. The entries "Source-only" and "Target-only" represent that the training uses either image only from the source and target domain.

Method	Classification Accuracy/%
Source-only	56.6
CycleGAN [27]	74.5
DRIT [16]	86.93
Ours	***90.4***
DANN [7]	77.4
DSN [4]	83.2
PixelDA [3]	**95.9**
Target-only	96.5

3.2.2 *Quantitative Evaluation*

Realism Evaluation The result of the realism evaluation comparison is shown in Table 3, in term of FID. As we can see from Table 3, on Cityscapes dataset, the FID of our method is 6.92, 7.55, 8.03, 12.09, 13.44, 15.68, and 15.75 lower than DRIT [16], UNIT [17], CycleGAN [27], DiscoGAN [12], DualGAN [25], Pix2PixHD [22], Pix2Pix [10], respectively. This means the generated images on Cityscapes dataset using our method is more realistic than using other methods. On Summer2Winter dataset, the FID of our method is 6.51, 7.52, 7.57, 8.71, and 8.72 lower than DRIT [16], UNIT [17], CycleGAN [27], DiscoGAN [12], DualGAN [25], respectively. It is obvious that the generated images Summer2Winter dataset using our method is more realistic than using other methods. In general, we find the FID of our method is larger than others. This suggests our method is better than others, and the result of our method are realistic.

Diversity Evaluation The result of the diversity evaluation comparison is shown in Table 4, in term of NDB. As we can see from Table 4, on Cityscapes dataset, the NDB of our method is 2.78, 3.90, 3.93, 7.36, 9.55, 96.49, and 96.56 lower than DRIT [16], UNIT [17], CycleGAN [27], DiscoGAN [12], DualGAN [25], Pix2PixHD [22], Pix2Pix [10], respectively. This means the generated images on Cityscapes dataset using our method is more diversified than using other methods. On Summer2Winter dataset, the NDB metric of our method is 2.14, 2.69, 3.05, 11.79, and 12.16 lower than DRIT [16], UNIT [17], CycleGAN [27], DiscoGAN [12], DualGAN [25], respectively. It is obvious that the generated images Summer2Winter dataset using our method is more diversified than using other methods. In general, we find the NDB of our method is larger than the others. This suggests our method is better than others, and the result of our method are diversified.

The result of the diversity evaluation comparison is shown in Table 5, in term of JSD. As we can see from Table 5, on Cityscapes dataset, the JSD of our method is 0.0292, 0.0300, 0.0590, 0.0680, 0.1330, 0.1419, and 0.1589 lower than DRIT [16], UNIT [17], CycleGAN [27], DiscoGAN [12], DualGAN [25], Pix2PixHD [22], Pix2Pix [10], respectively. This means the generated images on Cityscapes dataset using our method is more diversified than using other methods. On Summer2Winter dataset, the JSD metric of our method is 0.0297, 0.0298, 0.0587, 0.1128, and 0.1598 lower than DRIT [16], UNIT [17], CycleGAN [27], DiscoGAN [12], DualGAN [25], respectively. It is obvious that the generated images Summer2Winter dataset using our method is more diversified than using other methods. In general, we find the JSD of our method is larger than the others. This suggests our method is better than others, and the result of our method are diversified.

The result of the diversity evaluation comparison is shown in Table 6, in term of LPIPS. As we can see from Table 6, on Cityscapes dataset, the LPIPS metric of our method is 0.0096, 0.0215, 0.0094, 0.1922, 0.1997, 0.0213, 0.0149 higher than DRIT [16], UNIT [17], CycleGAN [27], DiscoGAN [12], DualGAN [25], Pix2PixHD [22], Pix2Pix [10], respectively. This means the generated images on Cityscapes dataset using our method is more diversified than using other methods. On Summer2Winter dataset, the LPIPS metric of our method is 0.0095, 0.0229, 0.0105, 0.1636, 0.1491 higher than DRIT [16], UNIT [17], CycleGAN [27], DiscoGAN [12], DualGAN [25], respectively. It is obvious that the generated images Summer2Winter dataset using our method is more diversified than using others. In general, we find the LPIPS metric of our method is larger than the others. This suggests our method is better than others, and the result of our method are diversified.

The result of diversity evaluation is shown in Table 7, in term of the reconstruction error. From Table 7, we find the reconstruction error of our method is smaller than the others. This shows that our method is better than other methods, and the result of image reconstruction using our method is best.

3.2.3 *Domain Adaptation*

It is demonstrated that the proposed image-to-image translation problem is beneficial to the adaptation of unsupervised domains. Following PixelDA [3], we use MNIST [15] to MNIST-M [7] for classification experiments. To evaluate our approach, we first translate the source image into the target domain. We then treat the generated labeled image as training data and train the classifier for each task in the target domain. For a fair comparison, we use the same architecture classifier as PixelDA [3]. We compare the proposed method with C [27] and DRIT [16] (based on our previous experiments to generate the most realistic images in the target domain) and three state-of-the-art domain adaptive algorithms: PixelDA [3], DANN [7], and DSN [4]. We report the classification accuracy on MNIST to MNIST-M, and the results are listed in Table 8. From Table 8, our results verify that our method can simulate different images in the target domain and improve performance in the target task, and is superior to other tasks.

4 Conclusion and Future Work

We proposed a novel image-to-image translation approach inspired by the human visual system. Moreover, we design an effective network loss to capture the pixel-level representations and human vision system information for verisimilar image-to-image translation. To enforce both structural and our model's consistency, we design Just-Noticeable-Difference loss based on a visual recognition task. The Just-Noticeable-Difference loss both guide the overall representation to be discriminative and enforce Just-Noticeable-Difference consistency before and after mapping between domains. We evaluate the proposed model through extensive qualitative and quantitative evaluation. In a wide variety of image-to-image tasks, we show diverse translation results with randomly sampled from existing images. We apply the proposed model to domain adaptation and show competitive performance when compared to the state-of-the-art on many datasets. In future research, we consider combining this idea of our method with video-to-video to get realistic generated video.

ACKNOWLEDGEMENTS

This work is supported by National Natural Science Foundation of China (61806198, 61533019, U1811463).

REFERENCES

[1] A. Anoosheh, T. Sattler, R. Timofte, M. Pollefeys, and L. Van Gool, 'Night-to-Day Image Translation for Retrieval-based Localization', *ArXiv e-prints*, (September 2018).
[2] S. Benaim and L. Wolf, 'One-Shot Unsupervised Cross Domain Translation', *ArXiv e-prints*, (June 2018).
[3] Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan, 'Unsupervised pixel-level domain adaptation with generative adversarial networks', in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (July 2017).

[4] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan, 'Domain separation networks', in *Advances in Neural Information Processing Systems 29*, eds., D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, 343–351, Curran Associates, Inc., (2016).

[5] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Scharwächter, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele, 'The cityscapes dataset', in *CVPR Workshop on the Future of Datasets in Vision*, volume 1, p. 3, (2015).

[6] Roland Fischer, Frances Griffin, Robert C. Archer, Stephen C. Zinsmeister, and Philip S. Jastram, 'Weber ratio in gustatory chemoreception; an indicator of systemic (drug) reactivity', *Nature*, **207**, 1049, (1965).

[7] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky, *Domain-Adversarial Training of Neural Networks*, 189–209, Springer International Publishing, Cham, 2017.

[8] Yunye Gong, Srikrishna Karanam, Ziyan Wu, Kuan-Chuan Peng, Jan Ernst, and Peter C. Doerschuk, 'Learning compositional visual concepts with mutual consistency', in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (June 2018).

[9] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter, 'Gans trained by a two time-scale update rule converge to a local nash equilibrium', in *Advances in Neural Information Processing Systems 30*, eds., I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 6626–6637, Curran Associates, Inc., (2017).

[10] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros, 'Image-to-image translation with conditional adversarial networks', in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (July 2017).

[11] L. J. Karam, N. G. Sadaka, R. Ferzli, and Z. A. Ivanovski, 'An efficient selective perceptual-based super-resolution estimator', *IEEE Transactions on Image Processing*, **20**(12), 3470–3482, (Dec 2011).

[12] Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim, 'Learning to discover cross-domain relations with generative adversarial networks', *CoRR*, **abs/1703.05192**, (2017).

[13] D. P Kingma and M. Welling, 'Auto-Encoding Variational Bayes', *ArXiv e-prints*, (December 2013).

[14] Diederik P. Kingma and Jimmy Ba, 'Adam: A method for stochastic optimization', *CoRR*, **abs/1412.6980**, (2014).

[15] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, 'Gradient-based learning applied to document recognition', *Proceedings of the IEEE*, **86**(11), 2278–2324, (Nov 1998).

[16] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang, 'Diverse image-to-image translation via disentangled representations', in *The European Conference on Computer Vision (ECCV)*, (September 2018).

[17] Ming-Yu Liu, Thomas Breuel, and Jan Kautz, 'Unsupervised image-to-image translation networks', in *Advances in Neural Information Processing Systems*, pp. 700–708, (2017).

[18] Xiao Liu, Shengchuan Zhang, Hong Liu, Xin Liu, and Rongrong Ji, 'Less is more: Unified model for unsupervised multi-domain image-to-image translation', *arXiv preprint arXiv:1805.10871*, (2018).

[19] Qi Mao, Hsin-Ying Lee, Hung-Yu Tseng, Siwei Ma, and Ming-Hsuan Yang, 'Mode Seeking Generative Adversarial Networks for Diverse Image Synthesis', in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019 , in press).

[20] Francesco Marra, Diego Gragnaniello, Davide Cozzolino, and Luisa Verdoliva, 'Detection of gan-generated fake images over social networks', in *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pp. 384–389. IEEE, (2018).

[21] Eitan Richardson and Yair Weiss, 'On gans and gmms', in *Proceedings of the 32Nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 5852–5863, USA, (2018). Curran Associates Inc.

[22] 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (June 2018).

[23] Arie Wisianto, Hidayatus Saniya, and Oki Gumilar, 'Integrating pipeline data management application and google maps dataset on web based gis application using open source technology sharp map and open layers', in *2010 8th International Pipeline Conference*, pp. 209–211. American Society of Mechanical Engineers, (2010).

[24] J. Wu, L. Li, W. Dong, G. Shi, W. Lin, and C. . J. Kuo, 'Enhanced just noticeable difference model for images with pattern complexity', *IEEE Transactions on Image Processing*, **26**(6), 2682–2693, (June 2017).

[25] Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong, 'Dualgan: Unsupervised dual learning for image-to-image translation', in *The IEEE International Conference on Computer Vision (ICCV)*, (Oct 2017).

[26] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang, 'The unreasonable effectiveness of deep features as a perceptual metric', in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (June 2018).

[27] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros, 'Unpaired image-to-image translation using cycle-consistent adversarial networks', in *The IEEE International Conference on Computer Vision (ICCV)*, (Oct 2017).

[28] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman, 'Toward multimodal image-to-image translation', in *Advances in Neural Information Processing Systems*, pp. 465–476, (2017).