IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 12, 2020, accepted June 28, 2020, date of publication August 26, 2020, date of current version September 3, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3007266

DuCaGAN: Unified Dual Capsule Generative Adversarial Network for Unsupervised Image-to-Image Translation

GUIFANG SHAO[1,2], MENG HUANG[1,2], FENGQIANG GAO[1,2,3], TUNDONG LIU[1,2], AND LIDUAN LI[1,2]

[1]Department of Automation, Xiamen University, Xiamen 361005, China
[2]Xiamen Key Laboratory of Big Data Intelligent Analysis and Decision, Xiamen 361005, China
[3]School of Information Science and Technology, Xiamen University Tan Kah Kee College, Zhangzhou, China

Corresponding author: Guifang Shao (gfshao@xmu.edu.cn)

This work was supported in part by the Industry and University Cooperation Project of Fujian Province under Grant 2018H6018, in part by the Science and Technology Project of Xiamen under Grant 3502Z20183004, and in part by the Program for Young Excellent Talents in University of Fujian Province under Grant 201847.

ABSTRACT With the appearance of Generative Adversarial Network (GAN), image-to-image translation based on a new unified framework has attracted growing interests. As a new technique, it can generate synthesizing images for various requirements in both computer vision and image processing. However, the cycle consistent structure adopted in some common models, such as cycle generative adversarial network (CycleGAN), is usually unable to learn more abundant image features. In this work, we developed a novel model based on GAN, named as dual capsule generative adversarial network (DuCaGAN), by utilizing the distinctive characteristic of view angle invariance and rotation equivariance in capsule network. Firstly, two capsule networks were introduced into the traditional CycleGAN model as discriminators to form our proposed model with six agents. To improve the feature capturing performance, we modified the full objective by combining the margin loss and the original adversarial loss. Furthermore, the Routing Algorithm in the capsule network was optimized by changing its compression function. Finally, experimental results on conventional visual tasks with paired and unpaired datasets demonstrated the superiority and effectiveness of the proposed approach compared to both deep convolutional generative adversarial network (DCGAN) and CycleGAN methods. More importantly, the proposed DuCaGAN was applied for the first time to augment the surface defect data from the real industrial field, and exhibited better performance than those methods available.

INDEX TERMS Image translation, generative adversarial network, capsule network, adversarial loss, data augmentation.

I. INTRODUCTION

Image-to-image translation, mapping an image from one domain to another, can resolve many problems in computer vision and image processing, such as texture synthesis, image super-resolution, image segmentation, style transfer, season transfer, and data augmentation [1]. For example, due to the time-consumption and high cost for creating a large amount of paired data for autonomous driving is time-consuming and costly, the image-to-image translation method is used to enrich the dataset of the autonomous driving scenes by synthesizing various street scene images to improve the learning ability [2], [3]. Especially, image augmentation applied to surface defect data from the real industrial field can also be formalized as the image translation problem [4], [5].

Actually, due to the multiple sub-phases and various devices used in a complete industrial process, there may exist different surface defects containing limited feature information on one product. Especially, these surface defects on the product appear occasionally and result in the rare defect samples. To enrich the defect sample images, the traditional image processing methods, such as copy, rotating, and cropping, are employed, while they can not display the defect features correctly. Therefore, insufficient sample sizes and sample class imbalances have become an urgent problem to be solved for the defect data in the real industrial process [4], [5]. Although some previous image translation methods had not focused on solving this problem, we can transfer it

The associate editor coordinating the review of this manuscript and approving it for publication was Orazio Gambino.

into an unsupervised image translation problem by respectively modeling samples as the normal domain and the defect domain, respectively. This transferring can not only augment the type of defect samples, but also can increase the number of image samples containing defect features. As a consequence, the accuracy of surface defect detection in industrial scenes can be improved.

In the past decades, various image translation methods have been proposed to address the aforementioned computer vision problems including image translation [6]–[12]. Some of them are put forward as a unified framework [2], and others are used to solve the shortage of industrial surface defect data and maintain sample imbalance [4]. With the development of the neural network, deep learning-based approaches have been introduced to solve some tasks like semantic segmentation, image coloring, image reconstruction, image super-resolution, data augmentation [1], [13], [14] and so on. These methods translate an image into another scene image by learning image features, including the deep multi-modal fusion network (DMFNet) [15], the fully convolutional neural network (FCN) [10], the common multi-scale convolutional architecture [16], and the deep feature fusion network (DFFNet) [17]. Although these works have made great progress in the single task of image translation, they fail to serve as a unified framework, especially in the application of image translation on industrial defect data augmentation. Therefore, an investigation of adversarial learning-based approach is of great significance to lots of computer vision tasks.

In recent five years, there are a multi-modal generative approach [18], the generative adversarial network (GAN) [19], and the dual conditional generative adversarial network [20] in the generative model. More importantly, the conditional generative adversarial network (CGAN) [11] has been put forward as a unified architecture to address the image-to-image translation problems mentioned above. Also, Phillip Isola *et al.* proposed a Pix2Pix model [2] by utilizing CGAN [11] to deal with the supervised learning problem. However, it is difficult to obtain enough complicated scene images with the corresponding label. Therefore, some researchers focused on the unsupervised image translation problem. For example, Schwing *et al.* put forward the DualGAN [21] based on the dual learning mechanism [22] and Taeksoo Kim *et al.* built the DiscoGAN [23] by integrating cross-domain relationships. Meanwhile, Jun-Yan Zhu *et al.* designed the cycle generative adversarial network (CycleGAN) [12] by introducing a cycle-consistent structure to achieve closed-loop interaction of information. These methods have made great progress on unsupervised image translation. However, little attention was paid on the image translation of industrial surface defect data to augment defect images from the real industry field.

Besides, there still exist two main problems for unsupervised image translation. On one hand, the image contents mapping precision between two image collections needs to be further improved to learn image data distribution accurately. On the other hand, the authenticity of translated images has a great influence on capturing structural information and global features adequately. It can be better applied to the real world by achieving high quality translated images. For instance, the application of the new unsupervised image translation method to increase the autopilot image samples not only can save costs of data preparation, but also can improve the accuracy of the model prediction. Moreover, this approach is beneficial to solving the problem of insufficient sample sizes and sample imbalances on surface defect data from the real industry field. All these made the unsupervised image translation become a challenge. Therefore, it is necessary to propose a more effective and unified unsupervised image translation framework to address the above problems, especially for augmenting industrial surface defect samples.

In this work, we fully utilized the superior competition mechanism of multi-agent [24] and the capsule network [25] in the GAN model to solve the unsupervised image-to-image translation problem. First of all, to generate images with richer global and local features including sample defect features in the industrial field, we introduced the capsule network as a discriminator into the GAN framework. Then, inspired by the competition mechanism of multi-agent, we designed a novel model with two generators and four discriminators instead of those traditional models [12], [23], [24]. Here, two capsule networks were employed to capture the distribution of the image domain more accurately and to solve the bidirectional image-to-image translation. Also, due to the introduction of the capsule network, the combination of adversarial cycle loss from CycleGAN [12] with the margin loss [25] yielded our full objective function, named as the dual capsule generative adversarial network (DuCaGAN) by us. Our method can generate more realistic images including some features from the target domain and learn the mapping of different domains more accurately. Especially, some features from defect data collected in the industry can be effectively learned in the framework of our unsupervised image translation. The code is available at https://github.com/linxi159/ DualCapsuleGAN.

The main contributions of the proposed model can be listed as follows:

(i) We developed a novel generative adversarial model to generate images with richer details and structural features for the unsupervised image translation.
(ii) The proposed method was applied for the first time to generate industrial surface defect samples containing rich defect features for data augmentation in the real industrial field.
(iii) The capsule network is introduced as the main discriminator of multi-agent competition mechanism for the unsupervised image-to-image translation.
(iv) The effectiveness of capsule networks was experimentally demonstrated through the minor optimization of routing algorithms in large-scale image discrimination.
(v) Finally, the proposed DuCaGAN was effectively evaluated on a large number of datasets.

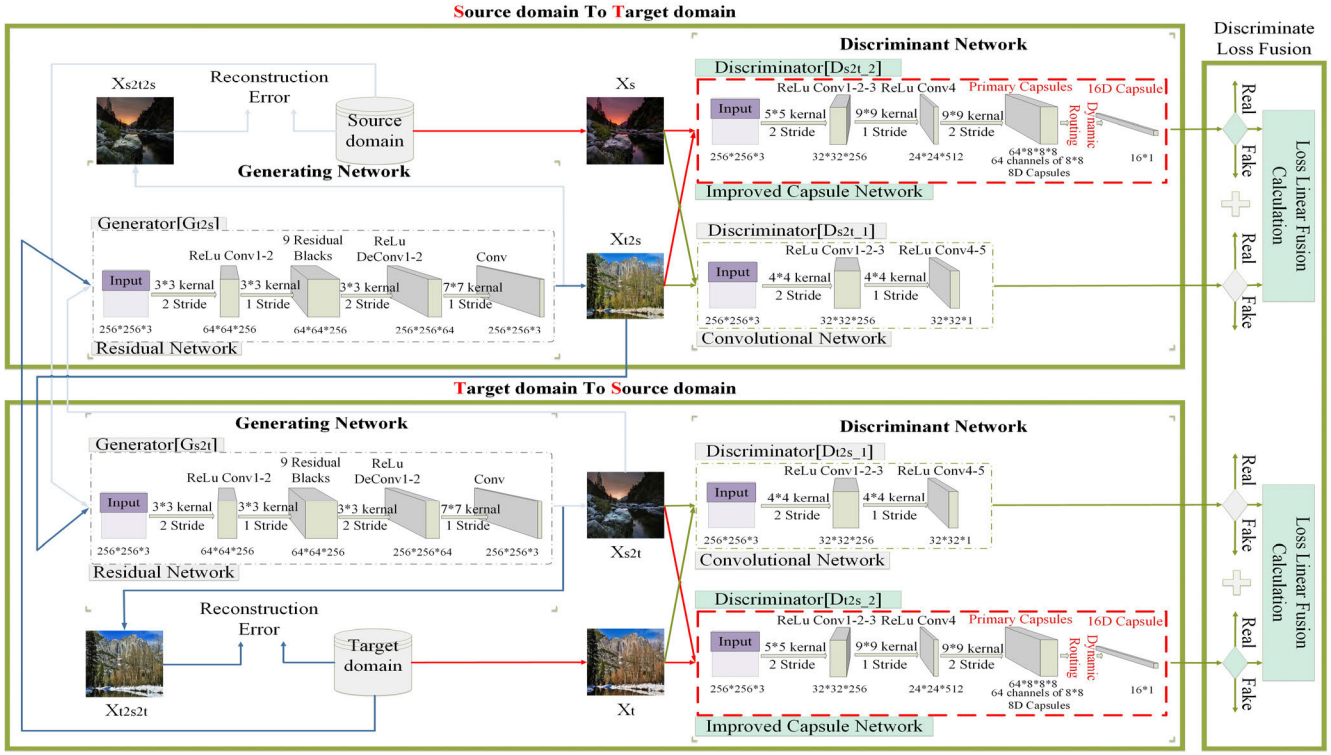

FIGURE 1. The framework of our proposed DuCaGAN model and the Capsule Network introduced in red dashed line.

II. METHOD

To address the image-to-image translation between two different domains, we developed a DuCaGAN model. This model based on the existing framework of CycleGAN [12]. optimizes the loss function and network structure Two capsule networks were introduced as the discriminators in our model to learn more detailed features, such as geometric features. By combining the margin loss function of capsule network [25], we also modified the previous loss function of GAN to improve the entire full objective and stabilize the training procedure. Also, the routing algorithm was optimized to avoid the inherent algorithmic flaws of the capsule network.

A. NETWORK ARCHITECTURE

As shown in Fig. 1, the framework of the proposed DuCaGAN model includes two modules. One of them is from the source domain to the target domain and the other is from the source domain to the target domain. In this model, there are four discriminators D and two generators G. Therefore, each module consists of two discriminators and one generator. It can be seen from Fig. 1 that the discriminant network in each module includes two discriminators D_{s2t-1} and D_{s2t-2}. They can distinguish source domain images X_s and the resulting images $G_{t2s}(x_t)$ from the source domain to the target domain. Here, the discriminator D_{s2t-1} is served by a convolution neural network including 3 convolution layers with stride-2 and 2 convolution layers with stride-1. The discriminator D_{s2t-2} is operated by our introduced capsule network composed of 3 convolution layers with stride-2, 1 convolution layers with stride-1, 1 primary capsule layer, and 16D capsule layer [25]. Additionally, the structure of Convolution-BatchNorm-LeakyRelu is used as the convolution calculation form of D_{s2t-1} and D_{s2t-2} [26]. Vice verse, a similar structure is used for the target domain to the source domain.

Given a generative network, it should contain one generator G_{s2t} or G_{t2s} in each module to generate more realistic images. Meanwhile, the generator mainly utilizes the residual network structure consisting of 2 convolution layers with stride-2 for down-sampling, 9 residual blocks [27] for training 256×256 images, 2 transposed convolution layers with stride-2 for up-sampling and 1 convolution layers with stride-1. The convolutional computation structure of the generator is from the Convolution-BatchNorm-Relu and Deconvolution-BatchNorm-Relu forms [26].

B. MODEL FORMULATION

A full objective usually consists of adversarial loss and reconstruction loss. The former aims to make the data distribution of generated images as similar as possible to that of the target domain and to generate a more realistic image. The latter is mainly used to prevent confusion and misplacement of the mapping relations between the source domain and the target domain. To learn the mapping relations between source domain X_s and target domain X_t more accurately,

we modified the full objective by introducing two new discriminators D_{s2t-2}, D_{t2s-2}. The modified full objective is displayed

$$\begin{aligned} &L_{G_{t2s},G_{s2t},D_{s2t-1},D_{s2t-2},D_{t2s-1},D_{t2s-2}} \\ &= L_{DuCaGAN}(G_{t2s}, D_{s2t-1}, D_{s2t-2}, X_s, X_t) \\ &\quad + L_{DuCaGAN}(G_{s2t}, D_{t2s-1}, D_{t2s-2}, X_t, X_s) \\ &\quad + \lambda_{rec} L_{rec}(G_{t2s}, G_{s2t}). \end{aligned} \tag{1}$$

$$\begin{aligned} &L_{DuCaGAN}(G, D_1, D_2, X, Y) \\ &= L_{DuCaGAN-1}(G, D_1, X, Y) + L_{DuCaGAN-2}(G, D_2, X, Y). \end{aligned} \tag{2}$$

where λ_{rec} is a hyper-parameter that shows the relative importance of reconstruction loss in the full objective. Here, $\lambda_{rec} = 10$.

$$\begin{aligned} &L_{DuCaGAN-k}(G, D_k, X, Y) \\ &\quad = E_{Y\sim p_{data}(Y)}[\log(D_k(Y))] \\ &\qquad + E_{X\sim p_{data}(X)}[\log(1 - D_k(G(X)))] \\ &\qquad + \lambda_k E_{Y\sim p_{data}(Y)}[-L_M(D_k(Y), T = 1)] \\ &\qquad + \lambda_k E_{X\sim p_{data}(X)}[-L_M(D_k(G(X)), T = 0)]. \end{aligned} \tag{3}$$

To improve the full objective and to avoid instability of training, we combined the margin loss [28], [29] and the original adversarial loss. The objective function is defined as Eq. 3, where λ_k are hyper-parameters that represent the relative importance of the margin loss in the improved adversarial loss compared to the original adversarial loss. Then, the margin loss we introduced to prevent training instability and mode collapse can be expressed as [21]:

$$v_k = CapsuleD(x_k). \tag{4}$$

$$\begin{aligned} L_M = &\sum_{k=1}^{K} T_k \max(0, m^+ - \|v_k\|)^2 \\ &+ \lambda(1 - T_k)\max(0, \|v_k\| - m^-)^2. \end{aligned} \tag{5}$$

in which, v_k refers to the output vector of the last layer in the capsule discriminant network and x_k is the input of the model in our framework. Here, G attempts to generate some samples G(x) that fuse domain Y features to the greatest extent. D_1 and D_2 try to make a distinction between translated images G(x) and real images Y. Then, G competes with D_1 and D_2 to achieve a Nash equilibrium. For example, more detailed information is stated as follows: the aim of the module from the source domain to the target domain is Eq. 6 and another is Eq. 7.

$$\begin{aligned} &\min_{G_{t2s}} \max_{D_{s2t-1},D_{s2t-2}} \\ &L_{DuCaGAN}(G_{t2s}, D_{s2t-1}, D_{s2t-2}, X_s, X_t) \end{aligned} \tag{6}$$

$$\begin{aligned} &\min_{G_{s2t}} \max_{D_{t2s-1},D_{t2s-2}} \\ &L_{DuCaGAN}(G_{s2t}, D_{t2s-1}, D_{t2s-2}, X_t, X_s) \end{aligned} \tag{7}$$

To make the authenticity of the image generated by the generator closer to the image of the target domain, it is necessary to minimize the adversarial loss (see Eq. 2). However, it is not guaranteed that the features learned in the target domain can be applied to other images in the source domain even if the adversarial loss is minimized. Therefore, during the training of the generator, we not only consider the adversarial loss, but also apply the cycle consistency loss L_{rec} [12], [23] to mitigate this problem.

$$\begin{aligned} L_{rec} &= (G_{t2s}, G_{s2t}) \\ &= E_{X_s\sim p_{data}(X_s)}[\|G_{t2s}(G_{s2t}(X_s)) - X_s\|_1] \\ &\quad + E_{X_t\sim p_{data}(X_t)}[\|G_{s2t}(G_{t2s}(X_t)) - X_t\|_1] \end{aligned} \tag{8}$$

C. ROUTING ALGORITHM OPTIMIZATION

For the capsule network [25] used in our novel model, its dynamic routing algorithm may cause wide controversy [30]. Therefore, some pioneering works have been done in two aspects. For instance, the vector is replaced with the matrix as a capsule structure [31] and the inter-capsule routing strategy is formalized into an optimization problem [30]. The former employed EM routing to learn the relationship of an entity and the pose in the matrix capsule structure instead of normal compression functions and the latter modified the compression function to optimize the routing algorithm but failed to perform compression operation during each iteration. The contribution of activation should be calculated at each iteration because the result of the compression function represents the activation probability of a higher-level capsule.

To solve the aforementioned problems and to avoid that the change of large values covers the change of small values, we try to modify the compression function inspired by some pioneering works of capsule network improvements [30], [31]. Our compression function is:

$$\omega_j = \frac{\sum_i c_{ij}\mu^*_{j|i}}{1 + \max_k \left\| \sum_i c_{ik}\mu^*_{k|i} \right\|} \tag{9}$$

where the c_{ij} refers to the coupling coefficient and can be determined during the iterative process of the dynamic routing algorithm. This coefficient indicates the tendency of low-level capsules i to high-level capsules j, and the higher the coefficient, the greater the tendency. The $\mu^*_{j|i}$ is the prediction vector of the capsule network and can be is obtained by multiplying the output of the last capsule layer and weight matrix.

D. THEORETICAL ANALYSIS

Based on Eq. 1-9, the theoretical analysis of the proposed DuCaGAN algorithm for image translation are given in this subsection. The definition of an optimal generator is expressed as follows:

$$G^* = \arg\min_G Div(P_{G(X)}, P_Y) \tag{10}$$

where X is the source domain data, Y is the target domain data, $P_{G(X)}$ is the generated data distribution in the original domain, and P_Y is the actual distribution in the target domain.

First, we obtained X, and use the generator to get G(x), and got $P_{G(X)}$ through the neural network. The goal is to minimize

the difference between $P_{G(X)}$ and P_Y. The optimization goal of the proposed method is given in Eqs. 6 and 7. Because every sample datum x comes from either the training set or the generated data, we defined the loss contribution of any sample x to the discriminator, as shown in Eq. 11.

$$CTB\_Dis = CTB\_D_1 + CTB\_D_2 \tag{11}$$

$$CTB\_D_1 = -P_Y(x)\log D_1(x) - P_{G(X)}(x)\log[1 - D_1(x)] \tag{12}$$

$$CTB\_D_2 = CTB\_D_1 - \lambda P_Y(x) L_M^{T=1}[Capsule\, D_2(x)] - \lambda P_{G(X)}(x) L_M^{T=0}[Capsule\, D_2(x)] \tag{13}$$

where CTB_Dis is the contribution of any sample x to all discriminators; CTB_D_1 is the contribution of any sample x to discriminator D_1; and CTB_D_2 is the contribution of any sample x to discriminator D_2.

Theorem 1: In the translation process between the source domain and the target domain, the minimax game achieves a Nash equilibrium when $P_{G(X)} = P_Y$.

Proof: From Eq. 12, we can see that the fixed generator G and the target domain data Y are regarded as constants, and then the partial derivative of D_1 is obtained, Dif = $-P_Y(x)/D_1(x) + P_{G(X)}(x)/(1 - D_1(x))$. The discriminator's optimal loss value is obtained when the derivative is 0, so as to obtain the optimal discriminator $D_1^*(x) = P_Y(x)/(P_Y(x) + P_{G(X)}(x))$. Finally, $D_1^*(x)$ is put to the D_1 loss function.

$$\begin{aligned}
&V(G, D_1)\\
&= E_{x\sim P_Y}[\log \frac{P_Y(x)}{P_Y(x) + P_{G(X)}(x)}]\\
&\quad + E_{x\sim P_G}[\log \frac{P_{G(x)}(x)}{P_Y(x) + P_{G(X)}(x)}]\\
&= \int_x P_Y(x)\log \frac{P_Y(x)}{P_Y(x) + P_{G(X)}(x)}dx\\
&\quad + \int_x P_{G(x)}(x)\log \frac{P_{G(x)}(x)}{P_Y(x) + P_{G(X)}(x)}dx\\
&= -2\log 2 + \int_x P_Y(x)\log \frac{P_Y(x)}{(P_Y(x) + P_{G(X)}(x))/2}dx\\
&\quad + \int_x P_{G(x)}(x)\log \frac{P_{G(x)}(x)}{(P_Y(x) + P_{G(X)}(x))/2}dx\\
&= -2\log 2 + KL(P_Y \left\| \frac{P_Y + P_{G(x)}}{2}\right.)\\
&\quad + KL(P_{G(x)} \left\| \frac{P_Y + P_{G(x)}}{2}\right.)\\
&= -2\log 2 + 2\cdot JSD(P_Y \| P_{G(x)})
\end{aligned} \tag{14}$$

According to the KL divergence property, KL(p || q) >= 0. If and only if p = q, the equality sign holds. Therefore, when the result of Eq. 14 is minimized, $P_Y = (P_Y + P_{G(x)})/ = P_{G(x)}$. That is, when $P_Y = P_{G(x)}$, the Nash equilibrium is obtained. The proof is completed.

From the derivation result of Eq. 14, one may see that training a discriminator is to maximize the JS divergence between the actual data and the sampled data of the generator. Therefore, the goal of the discriminator in the proposed method is expressed as follows:

$$G^{\wedge} = \arg\min_G \max_D V(G, D_1, D_2, X, Y) \tag{15}$$

in which one finds the generator G and fix it, and then looks for the discriminator D to maximize the JS divergence of Y and G(x). That is the discriminator can distinguish the actual distribution and the generated distribution. At the same time, the generator G is adjusted to minimize the difference value (JS divergence) between G(x) and Y. That is, the difference between the distribution generated by the generator and the actual distribution is minimized.

Theoretically, it is more difficult to find the nash equilibrium in the actual training process than to optimize the objective function. In other words, it is difficult to make P_Y equal to $P_{G(x)}$. Therefore, we improved the overall discriminative ability of the proposed method by adding a capsule network as an additional discriminator D_2 to achieve a nash equilibrium. It is seen from Eq. 13 that CTB_D_2 denotes the effect of the capsule network in the proposed DuCaGAN.

E. IMPROVED TRAINING PROCEDURE

In this work, we proposed a high-level training procedure for the DuCaGAN model in Algorithm 1.

Algorithm 1 DuCaGAN Algorithm

1:**arguments:**
  Generator: G_{s2t}, G_{t2s}.
  Discriminator: D_{s2t-1}, D_{s2t-2}, D_{t2s-1}, D_{t2s-2}.
2:**initialization:**
  The networks and other pertinent hyper-paraeters
3: **for** number of epochs **do**
4: **for** number of training steps **do**
5: Sample minibatch of n samples $X_s = \{x_s^{(1)}, \ldots, x_s^{(n)}\}$ from $p_{data}(x_s)$.
6: Sample minibatch of n samples $X_s = \{x_s^{(1)}, \ldots, x_s^{(n)}\}$ from $p_{data}(x_t)$.
7: Update generator G_{t2s}: by minimizing generating loss g_{s2t}.
8: Update generator G_{s2t}: by minimizing generating loss g_{t2s}.
9: Update discriminator D_{s2t-1}, D_{s2t-2}: by minimizing discriminating loss d_{s2t}.
10: Update discriminator D_{t2s-1}, D_{t2s-2}: by minimizing discriminating loss d_{t2s}.
11: **end for**
12: **end for**

In the previous CycleGAN model, there are four networks, *i.e.*, the two generators, and two discriminators are optimized as a whole. There are six networks and corresponding hyper-parameters to be initialized in the DuCaGAN model. Besides, to learn the distribution of different domains more accurately, the two generators and four discriminators in the proposed model are optimized separately and independently. In step 2, the two new capsule networks D_{s2t-2} and D_{t2s-2} need to be initialized in the proposed framework. In step 7 to 10,

TABLE 1. The detailed information on 9 datasets.

NO.	Datasets	Usages	Instances	Size	DSC.	Add.
1	Cityscapes	unpaired	2975,2975	256*256	The city street view image and label	[32]
2	Sketch2photo	unpaired	995,995	256*256	The sketch and actual image of human face	[33]
3	Day2night	unpaired	90,90	512*512	Natural images at different moments in the same scene	[34]
4	Oil2chinese	unpaired	1177,1175	512*512	Oil and Chinese paintings of different artistic styles	[34]
5	Summer2winter	unpaired	1231,962	256*256	The natural scene images of different seasons	[12]
6	Ukiyoe2photo	unpaired	562,6287	256*256	The images of ukiyoe style and natural scenes	[12]
7	Vangogh2photo	unpaired	400,6287	256*256	The images of vangogh style and natural scenes	[12]
8	Surface defect data	unpaired	1044,924	256*256	The defective and normal images in aluminum profile datasets	--
9	DAGM 2007	unpaired	720,720	512*512	The defective and normal images in industrial optical datasets	--

G_{t2s}, G_{s2t}, and D_{s2t-1}, D_{s2t-2}, and D_{t2s-1}, D_{t2s-2} are updated independently in each turn.

III. EXPERIMENT AND RESULTS

A. EXPERIMENT SETTING

To evaluate the efficiency of the proposed DuCaGAN, experiments on various tasks are conducted. Here, 9 datasets are used through the unpaired usages, as shown in Table 1. The first dataset is an image segmentation dataset Cityscapes, designed to evaluate the performance of visual algorithms in urban scene semantic understanding. The Sketch2photo and Day2night are trained in an unpaired manner. The Oil2chinese, Ukiyoe2photo and Vangogh2photo are unpaired datasets that represent the style and scene translations between different art painting images or art painting and natural images. The Summer2winter is an unpaired dataset containing summer and winter scenes. The last two datasets are the surface defect data of aluminum profile acquired from an actual aluminum profile production in the industrial field and the industrial optical data. They are also the unpaired dataset that contains some defect images and normal images. These aluminum profile defect data and DAGM 2007 are provided by Tianchi Big Data Zhongzhi Platform-Aliyun Tianchi and 29th Annual Symposium of the German Association for Pattern Recognition respectively. All images in the above 9 datasets, which are three-channel color images, have been unpairing for the training process of all experiments.

Furthermore, to verify the superiority of DuCaGAN, we compared it with DCGAN [35], CycleGAN [12], and DD-CycleGAN [7]. Here, **DCGAN** uses adversarial loss to learn the mapping relations between two different domains. **CycleGAN** uses adversarial loss and introduces cycle consistency loss to regularize the mapping. Besides, **DD-CycleGAN** introduces double discriminators in CycleGAN to generate images.

Also, the parameter values used in our model are listed in Table 2 according to some methods for stochastic optimization [36]. The learning rate 0.0002 is used to train the model for the first 100 epochs and the linearly decaying strategy is utilized over the next 100 epochs. For parameters λ_1 and λ_2 in Eq. 3, we set λ_1 to be 0, 0.5 and 1 to conduct comparative experiments respectively, as well as λ_2, and finally obtained the suitable values. In the Eq. 5, the parameter λ is set to be 0.5 based on the marginal loss calculation [25]. The value of λ_{rec} is set to avoid dramatic changes in large values [37], [38].

TABLE 2. Parameter values.

parameter	β_1	β_2	Batch size	m^+	m^-
value	0.5	0.999	1	0.9	0.1
parameter	λ_1	λ_2	Learning rate	λ_{rec}	T_k
value	0.5	0.5	0.0002	10	0-1

Besides, the training images can be flipped, enlarged, cropped and rotated randomly. To avoid excessive oscillation of the model [20] and training instability [37], two types of generators and two types of discriminators were used to perform calculations respectively, which means that the weights were updated separately. All experiments were conducted on a single NVIDIA GeForce GTX 1080 GPU.

B. METRICS

Although the evaluation of the generated model is a quite difficult problem, we can use the semantic segmentation evaluation method to quantitatively evaluate the accuracy of generated images. As the FCN-score method [10] is usually utilized for pixel-level prediction, we adopted it to predict the output performance of the model trained in the Cityscapes dataset. Besides, four evaluation criteria were used to measure pixel accuracy and region intersection over union (IU). Let n_{ij} be the number of pixels in class i which is not predicted as class j and n_{cl} represents the number of different classes. The total pixel number in class i are $t_i = \sum_j n_{ij}$. Thus, four quantitative evaluation indicators can be calculated as follows [10].

The pixel accuracy (Per-pixel acc.) is defined as $\sum_i n_{ii} / \sum_i t_i$.

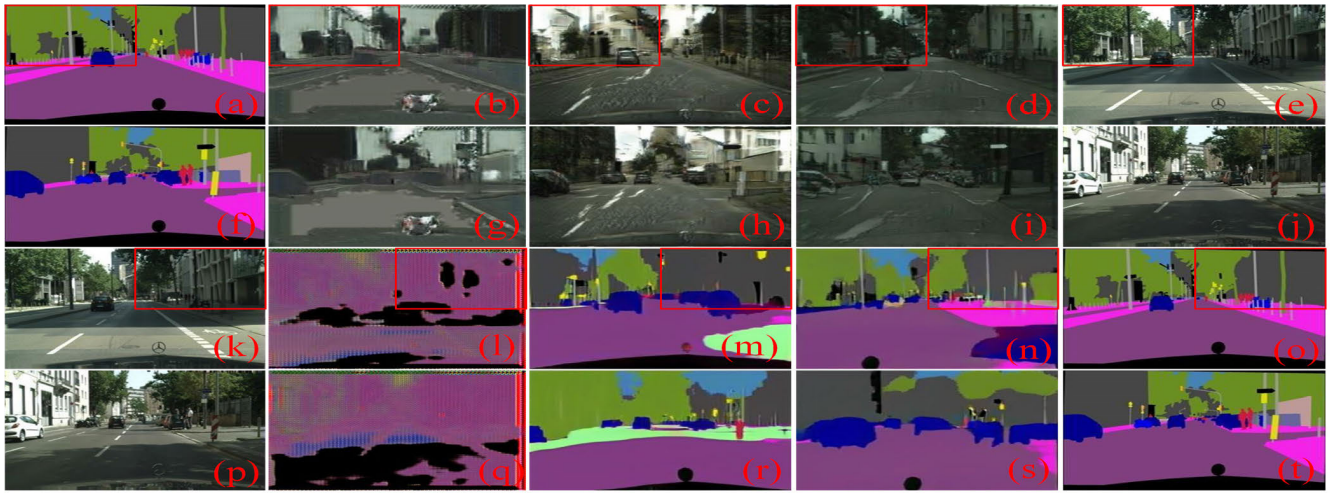

FIGURE 2. Results by different methods for labels ←→ photos translation in Cityscapes dataset.

The mean accuracy (Per-class acc.) is denoted as $(1/n_{cl}) \sum_i n_{ii}/t_i$.

The frequency weighted Intersection-Over-Union (Frequency weighted IOU) can be computed by $(\sum_k t_k)^{-1} \sum_i t_i n_{ii} /(t_i + \sum_j n_{ji} - n_{ii})$.

The mean class Intersection-Over-Union (ClassIOU) can be obtained by $(1/n_{cl}) \sum_i n_{ii}/ (t_i + \sum_j n_{ji} - n_{ii})$.

C. ANALYSIS OF CITYSCAPES DATA

1) QUALITATIVE COMPARISON WITH OTHER BENCHMARKS

To evaluate the performance of our model, comparative experiments with DCGAN, and CycleGAN on cityscapes dataset are conducted. Fig. 2 illustrates the translation results between the label domain to photo domain. If the inputs are images from the label domain, the original data are images form the photo domain. As shown in Fig. 2, DuCaGAN can learn more rich features of the target domain and preserve the structural information of the source domain well. In terms of the matching degree of structural features, labels, and the authenticity of generated images, DuCaGAN outperforms DCGAN and CycleGAN. Also, the generated images are more reasonable. As shown in Figs.2d, 2e, 2i, 2j, 2n, 2o, 2s, and 2t, the outputs of our model are maximally consistent with the contents of Original Data. On the contrary, the real image generated by DCGAN is not real enough due to the instability of the model as shown in Fig. 2b and 2g. The generated label is greatly distorted for DCGAN and the basic structure and color features are not captured well as shown in Figs. 2l and 2q. For CycleGAN, the serious problem is that the generated photos fail to accurately maintain consistency with the contents from the local object of a label, and vice versa. For example, some structural features of trees, buildings and ground have been misplaced during translation as shown in Figs. 2c, 2h, 2m, and 2r.

To further verify the effectiveness of our model, we magnify the red rectangle part in the first and third lines of Fig. 2, as shown in Fig. 3. From Figs. 3b to 3d, one can see that DuCaGAN generated more reasonable natural scene content, such as the label information of trees and houses. But DCGAN did not generate trees, and CycleGAN generated other objects instead of trees as the red circle displayed. Therefore, DuCaGAN can easily learn some structure, texture and color features of small objects depending on the powerful learning capability of the capsule network.

2) QUANTITATIVE COMPARISON WITH OTHER BASIC METHODS

To further evaluate the effectiveness of our method from a quantitative perspective, we mainly studied some different models and the FCN-score of these results.

Table 3 and Table 4 illustrate the statistical FCN-score of image translation results on the Cityscapes dataset in multiple experiments. From Table 3, one can see that DuCaGAN achieves a higher score than DCGAN and CycleGAN, which means that it learns richer local features of the target domain and also preserves the structural information of the source domain. Moreover, the Per-pixel acc. and Per-class acc. values of DuCaGAN are close to the Original Data, indicating our model possesses a higher recovery performance of generated images. Also, the higher Class IOU and Frequency weighted IOU values in the DuCaGAN demonstrate our model can generate a more accurate and reasonable structure.

The similar results are obtained for the cityscapes photos-to-labels task (see Table 4). All in all, our approach outperforms DCGAN and CycleGAN, and possesses closer results to Original Data.

3) ABLATION STUDY: ANALYSIS OF LOSS FUNCTION AND CAPSULE MODULE

The loss function is a unique part of different models resulting in various performance. Here, cycle loss referred to cycle consistent loss introduced by CycleGAN, original loss consists

Input DCGAN[35] CycleGAN[12] **DuCaGAN** Original Data

FIGURE 3. Detailed results in the first and third rows of Fig. 2 for labels ←→ photos translation.

TABLE 3. Fcn-scores for different methods on Cityscapes labels → photos.

Method	Frequency weighted IOU	Per-pixel acc.	Per-class acc.	Class IOU
DCGAN[35]	0.410 ± 0.006	0.522 ± 0.012	0.174 ± 0.007	0.115 ± 0.005
CycleGAN[12]	0.450 ± 0.037	0.539 ± 0.062	0.154 ± 0.012	0.106 ± 0.014
DuCaGAN	**0.543 ± 0.005**	**0.661 ± 0.015**	**0.232 ± 0.011**	**0.151 ± 0.002**
Original Data [10]	0.816 ± 0.002	0.892 ± 0.002	0.523 ± 0.004	0.443 ± 0.004

TABLE 4. Performance of photos → labels for different methods on Cityscapes.

Method	Frequency weighted IOU	Per-pixel acc.	Per-class acc.	Class IOU
DCGAN[35]	0.251 ± 0.122	0.356 ± 0.123	0.091 ± 0.034	0.049 ± 0.023
CycleGAN[12]	0.342 ± 0.008	0.441 ± 0.014	0.125 ± 0.019	0.075 ± 0.017
DuCaGAN	**0.375 ± 0.007**	**0.479 ± 0.014**	**0.167 ± 0.023**	**0.099 ± 0.014**
Original Data [10]	0.816 ± 0.002	0.892 ± 0.002	0.523 ± 0.004	0.443 ± 0.004

TABLE 5. Ablation study: FCN-scores for different variants of our method on Cityscapes labels → photos.

Variants	Frequency weighted IOU	Per-pixel acc.	Per-class acc.	Class IOU
CycleGAN + Cycle loss	0.321 ± 0.087	0.406 ± 0.128	0.119 ± 0.017	0.067 ± 0.017
CycleGAN + Original loss	0.450 ± 0.037	0.539 ± 0.062	0.154 ± 0.012	0.106 ± 0.014
DuCaGAN + Original loss	0.505 ± 0.073	0.608 ± 0.094	0.186 ± 0.025	0.131 ± 0.025
DuCaGAN + Optimal loss	**0.543 ± 0.005**	**0.661 ± 0.015**	**0.232 ± 0.011**	**0.151 ± 0.002**

TABLE 6. Ablation study: Classification performance of photos → labels for different variants on Cityscapes.

Variants	Frequency weighted IOU	Per-pixel acc.	Per-class acc.	Class IOU
CycleGAN + Cycle loss	0.121 ± 0.060	0.267 ± 0.054	0.072 ± 0.010	0.028 ± 0.011
CycleGAN + Original loss	0.342 ± 0.008	0.441 ± 0.014	0.125 ± 0.019	0.075 ± 0.017
DuCaGAN + Original loss	0.368 ± 0.041	0.466 ± 0.051	0.137 ± 0.038	0.079 ± 0.028
DuCaGAN + Optimal loss	**0.375 ± 0.007**	**0.479 ± 0.014**	**0.167 ± 0.023**	**0.099 ± 0.014**

of cycle loss and adversarial loss used in GAN, and we combined margin loss of capsule module with original loss as the optimal loss of DuCaGAN. We performed the ablation study of our full loss and framework to evaluate our loss function and capsule module, as Table 5 and Table 6 illustrated.

From the first and second row in Tables 5 and 6, it can be seen that the original loss performs better because CycleGAN not only used cycle loss, but also considered the loss of cycle consistency structure. As the third row of Tables 5 and 6 shown, DuCaGAN achieves a better result when we used the same loss function. This mainly attributes to the introduction

of dual capsule network structure in our model. What's more, an optimal loss as the full objective was utilized to further improve the performance of DuCaGAN. The last row in Table 5 and 6 also indicates the superiority of optimal loss compared with original loss.

From the first and second row in Tables 5 and 6, it can be seen that the original loss performs better because Cycle-GAN not only uses cycle loss, but also considers the loss of cycle consistency structure. As the third row of Tables 5 and 6 shown, DuCaGAN achieves a better result when we used the same loss function. This mainly attributes to the introduction of dual capsule network structure in our model. What's more, an optimal loss as the full objective is utilized to further improve the performance of DuCaGAN. The last row in Table 5 and 6 also indicates the superiority of optimal loss compared with the original loss.

4) PERCEPTUAL VALIDATION

Instead of quantitative analysis, we also introduce the perceptual realism to value our results for photos-to-labels translation in the Cityscapes dataset. Firstly, 10 images and their corresponding labels are randomly selected from the results of image translation achieved by different methods. Then, 25 participants are invited to compare these translated images with the labels. They give a score of 1 to 5 for each image sample according to their similarity judgment between the translated images and labels. The better the image quality is, the higher score will be.

From Fig. 4, one can see that the average score of our method is higher than that of CycleGAN and DCGAN. This also illustrates our method obtains higher image quality from the perspective of visual perception. Besides, we can find that the model collapse of DCGAN occurs in photos-to-labels translation from the results of DCGAN in Fig. 4(a). Especially, there is a higher standard deviation for perceptual validation of images translated by each method. This also means the man-made evaluating method of image quality has great instability. Therefore, it may only be used as an auxiliary approach to verify the performance of different methods.

D. ANALYSIS OF IMAGE TRANSLATION BASED ON CONTENTS

Fig. 5 displays the result obtained by different methods on the sketch2photo dataset. From Fig. 5, it can be seen that our method can capture more rich and specific details, so as to generate images that are closer to the Original Data compared with other methods. From the first and second rows in Fig. 5, we can see that DuCaGAN learns the whole brightness and color characteristics more accurately, and the generating images are closer to the Original Data. Additionally, as the third and fourth rows in Fig. 5 shown, DuCaGAN captures the face structural and contour features more clear than the other methods. Especially for some local features from the face, such as eyebrows and beards, the result of DuCaGAN is closer to real distribution. However, there is still a certain gap between the output of DuCaGAN and the Original Data for the chin.

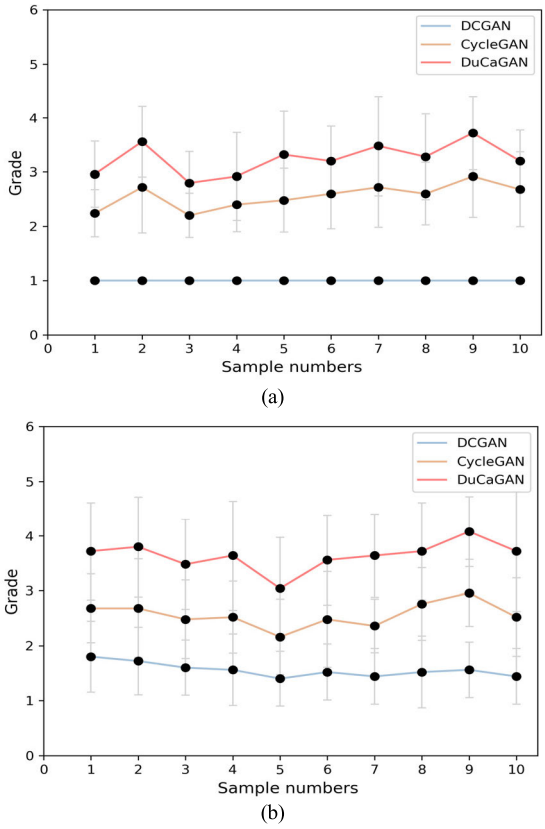

FIGURE 4. Results manually rated for image quality. The figure shows the rating results of images generated by different methods for photos ←→ labels translation on Cityscapes dataset.

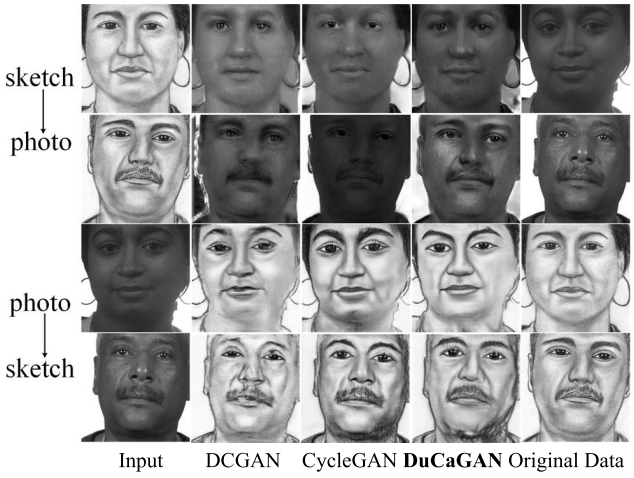

FIGURE 5. Results by different methods for sketch ←→ photo translation on sketch2photo dataset.

E. ANALIYSIS OF IMAGE QUALITY AND PERFORMANCE DURING MODEL ITERATIVE TRAINING

To further verify the detail feature capturing capacity of our proposed method, Table 7 illustrates the results at different iteration steps on the Day2night dataset. For the night to the day translation task, DuCaGAN can generate a more realistic

TABLE 7. Generated images by different methods for night → day translation at different iteration steps.

Method	5000 step	40,000 step	80,000 step	100,000 step
Input				
DCGAN[35]				
CycleGAN[12]				
DuCaGAN				
Original Data				

(a) Input (b) DCGAN[35] (c) CycleGAN[12] (d) **DuCaGAN** (e) Original Data

FIGURE 6. Detailed results of red rectangle area in Table 7.

image and get closer to the distribution of ground truth more quickly. However, DuCaGAN performs worse at 5000 step due to a large number of parameters to be optimized during the initial training. Besides, DuCaGAN can learn the primary features of the day (target domain) scene accurately and quickly compared to other methods with the iteration growing. Also, to prove the effectiveness of our proposed method clearly, we magnify the red rectangle area at 100,000 step of Table 7 as shown in Fig. 6. DuCaGAN can capture edge features more accurately, especially for brightness features, and the generating image is sharper. However, the images generated by DCGAN and CycleGAN are both too dark, illustrating that they did not capture the daylight feature.

Furthermore, we analyze the statistical loss values of the discriminant module for different methods on the day2night dataset in multiple experiments, as Fig. 7 illustrated. From Fig. 7, one can see that the discriminant loss of DuCaGAN is the middle one and less than the CycleGAN method after 100,000 iterations of training. Although the discriminant loss of DCGAN drops so fast even close to zero, it fails to effectively distinguish the generated image from the real one. The brightness features of the night scene are not well learned by DCGAN because of an over-fitting for this model. As the generated image at 5,000 step, 40,000 step, and 100,000 step shown, the model trained by DCGAN is unstable and the generated image is too deformed. We can conclude that DCGAN only learns some basic structural features with the iteration increasing. Meanwhile, CycleGAN also fails to capture the brightness features of the night scene during the initial training phase, but it performs better with the iteration growing. In contrast, DuCaGAN can learn the structure, color and brightness features of the target domain in a short iterative step. Moreover, the model trained by DuCaGAN becomes more stable as the training progresses.

Thus, as Fig. 7 illustrated, our method performs better than the other two methods during different iteration stages. This indicates that DuCaGAN can not only capture more detailed features of the target domain quickly, but also generate a more accurate distribution that approximates to the true one.

F. APPLICATION OF SEASON AND STYLE TRANSFER

Fig. 8 shows the generated result of three methods on the summer2winter Yosemite dataset. Obviously, for the task of summer-to-winter, DuCaGAN not only remains the structural information of the summer scene, but also effectively learns the characteristics of a large number of snow in winter scenes.

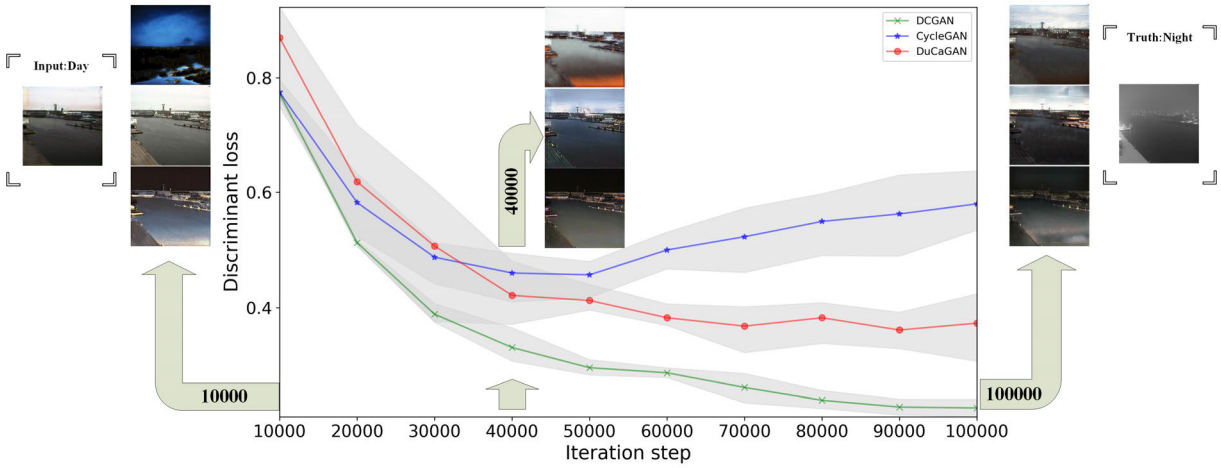

FIGURE 7. Results for day → night translation. The figure shows the discriminant loss of different methods and the generated image at the 5,000, 40,000 and 100, 000 iteration step. From top to bottom: DCGAN[35], CycleGAN[12] and DuCaGAN(our).

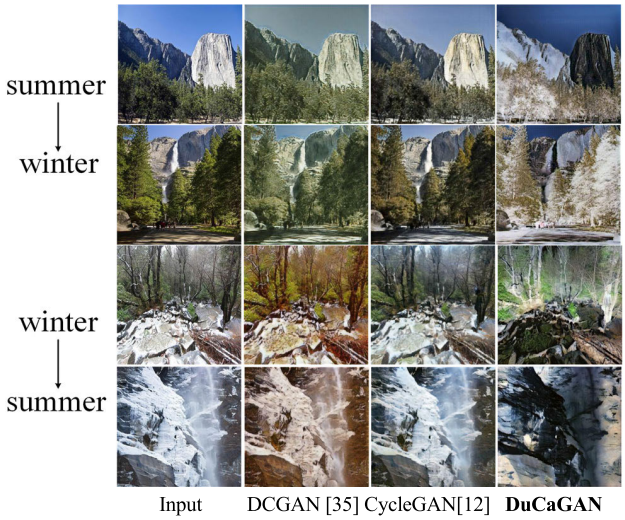

FIGURE 8. Results by different methods for summer ←→ winter Yosemite translation on summer2winter Yosemite dataset.

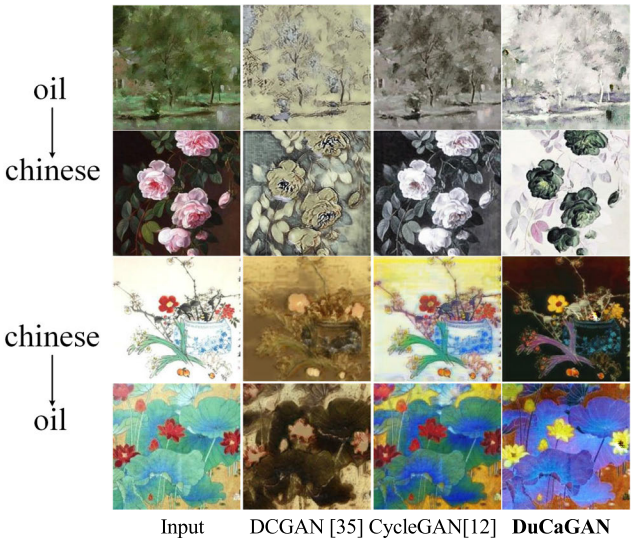

FIGURE 9. Results for oil painting ←→ Chinese painting translation.

For the task of winter-to-summer, DuCaGAN can effectively remove the snow feature of the winter scene. What's more, it finishes the knowledge transfer perfectly by capturing the color features of plants in the summer scene more clearly. However, the image color produced by DCGAN is not appropriate compared to the target domain, yet the color features of a local object captured by CycleGAN are not obvious enough as the third and fourth rows of Fig. 8 shown.

Figs 9, 10, and 11 display the image translation results between different artistic styles and natural scenes including oil2chinese painting, ukiyoe2photo and vangogh2photo. For the oil-to-chinese task of Fig. 9, DuCaGAN can capture the black and white characteristics, light color and ink texture of Chinese painting, and generate images that are more consistent with the distribution of the target domain. For the chinese-to-oil task of Fig. 9, DuCaGAN is able to learn more rich and diverse color features to produce more realistic images compared with other methods. As shown in Figs. 10 and 11, DuCaGAN can also perform style translation based on retaining the structural and textural information of the source domain. Compared to DCGAN and CycleGAN, it is more sensitive to some color, texture and artistic style features, especially for the color features. Moreover, for the task of vangogh to photo in Fig. 11, DCGAN has a model crash, resulting in the generated image not containing valid information.

In summary, DuCaGAN can generate an accurate distribution that is similar to the input image even though there is no corresponding label image in the target domain. In other words, it can produce a more convincing and realistic image.

G. ANALYSIS OF RECONSTRUCTED IMAGE

We made use of the reconstruction image $G_{t2s}(G_{s2t}(x))$ to analyze the effectiveness of our proposed method. That lets

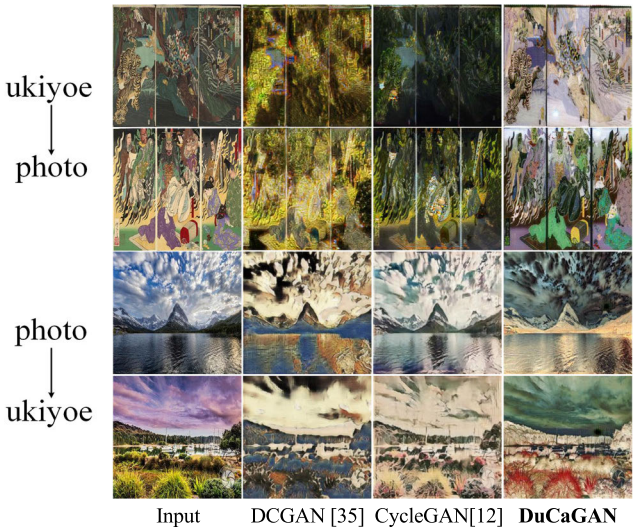

FIGURE 10. Results for ukiyoe ⟷ photo translation.

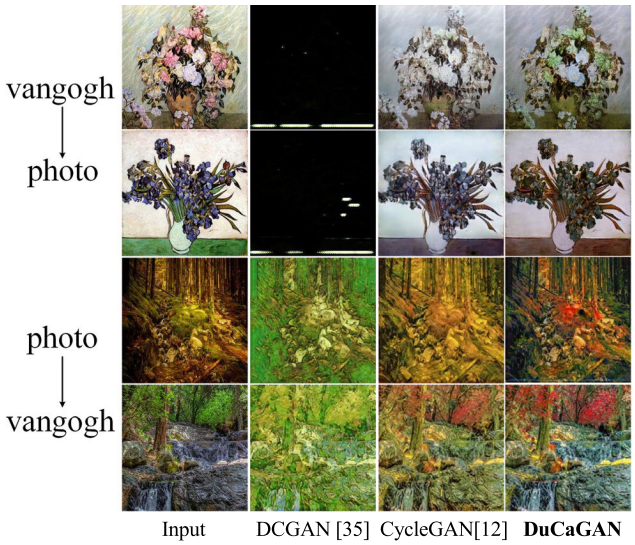

FIGURE 11. Results for vangogh ⟷ photo translation.

the generated image of source-to-target task become the input of target to source task. If the method is good enough, it will produce the same image as the input. We randomly selected four samples from Cityscapes, summer2winter Yosemite, oil2chinese and day2night datasets, and let their corresponding reconstructed images be input to get reverse reconstruction results as shown in Fig. 12. It can be seen that the reconstructed image generated by DuCaGAN is more similar to the original image compared to other models, which means that our model's generator performs better.

Owing to a large number of parameters in our model, it is easier to learn detailed features of texture, color and style and generate clearer images when considering the reconstruction information.

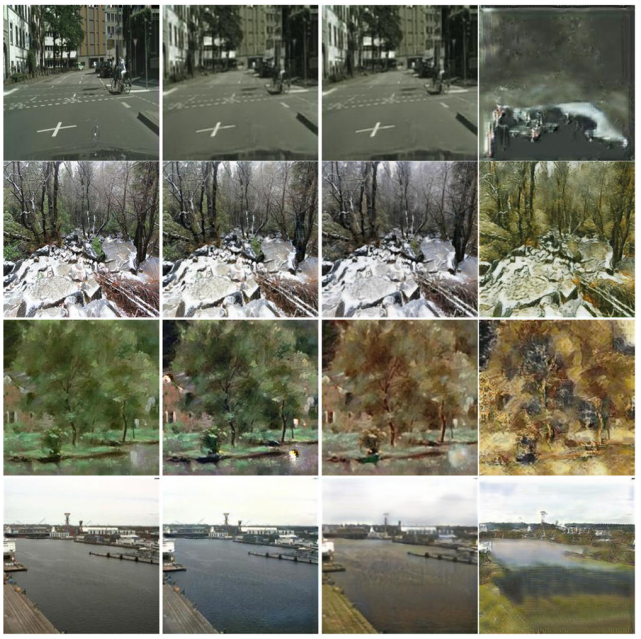

FIGURE 12. The input images x and the reconstructed images Gt2s(Gs2t(x)) obtained by different methods. Images are randomly select from top to bootom: labels → photos in Cityscapes dataset, winter → summer Yosemite dataset, oil → chinese dataset and day → night dataset.

H. PERCEPTUAL VALIDATION

Here, to evaluate our results, we perform the perceptual realism for summer-to-winter_Yosemite translation in the unpaired summer2winter Yosemite dataset. The experimental setting was similar to that of the perceptual validation in the paired data. The average score of our method is higher compared with CycleGAN and DCGAN, as Fig. 13 shown. This also indicates DuCaGAN can obtain a better analysis of image quality under the validation of visual perception and the winter or summer images generated by our method are more realistic than the other methods. Especially, there is a higher standard deviation for perceptual validation of images translated by each method in Fig. 13a and b. From the results, there is a great instability for the man-made evaluating method of image quality. Therefore, it may only be used as an auxiliary approach to verify the performance of different methods.

I. ANALYSIS OF ALUMINUM PROFILE DATA

1) ANALYSIS OF IMAGE QUALITY BASED ON DATA AUGMENTATION

To verify the performance of our method on the application of industrial data augmentation, experiments on the aluminum profile dataset obtained from the real industrial field were conducted. The normal and defective images produced by the industry are modeled as normal domain and defect domain respectively, and entered into different models.

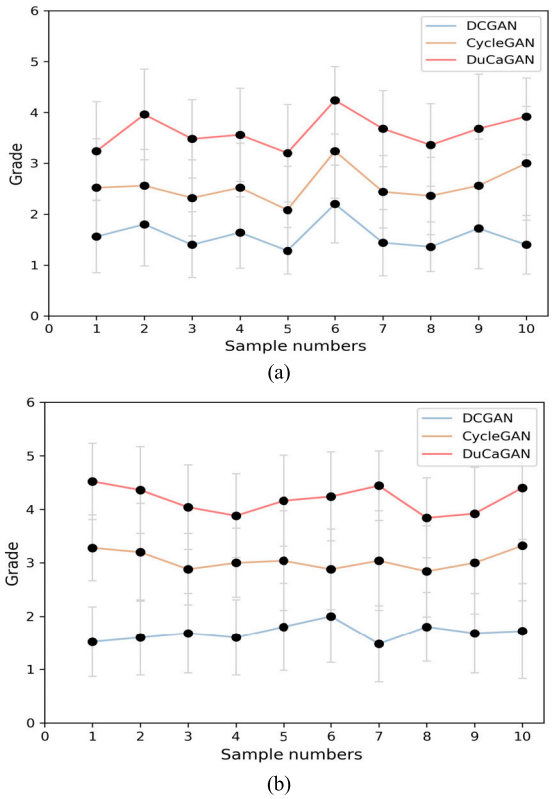

FIGURE 13. Results manually rated for image quality. The figure shows the rating results of images generated by different methods for summer ←→ winter Yosemite translation on summer2winter Yosemite dataset.

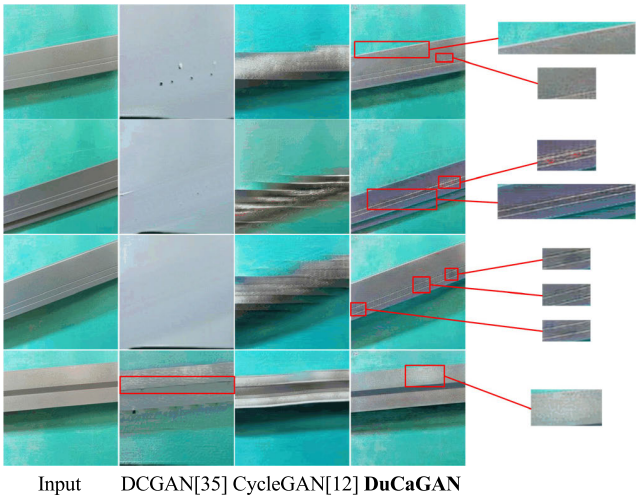

FIGURE 14. Results for normal images → defective images on aluminum profile dataset.

Fig. 14 illustrates the results of three different methods. We can see that the defective images generated by DCGAN are quite different from the input one, indicating that it fails to capture the effective structural features and appears model collapse. Moreover, the target object in the generated image of CycleGAN is deformed. For DuCaGAN, it not only retains the structural information of the original image, but also adds different defect features. Therefore, our method performs better than CycleGAN and DCGAN on the augmented image quality.

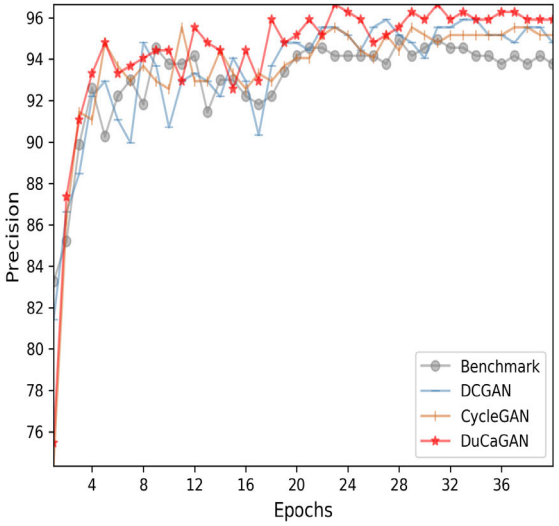

FIGURE 15. Accuracy of aluminum surface defect classification on defect data augmented by different models in aluminum profile dataset.

2) APPLICATION OF DATA AUGMENTATION

To further verify the effectiveness of our proposed method, we added the defective images augmented by different methods into the original aluminum profile dataset to form a new dataset, then analyzed the performance of aluminum surface defect classification. Fig. 15 shows the classified results of different methods, it is clear that DuCaGAN outperforms benchmark and other methods during model iterations. Moreover, the classification accuracy of our method becomes gradually stability after the 18th epoch and the highest value is reached at the 23rd and 31st epoch. However, DuCaGAN performs worse during the early stages of training because there are a large number of parameters to be optimized in our model to generate images containing more defect features. Besides, the classification performance of CycleGAN and benchmark starts to be gradually stable after the 21st epoch. The performance of DCGAN has been unstable throughout the training period because of the worse defect samples augmented by the collapsed model.

Therefore, we can conclude that our approach performs best on the augmentation and application of industrial data.

J. ANALYSIS OF INDUSTRIAL OPTICAL DATA

1) ANALYSIS OF IMAGE QUALITY FROM DIFFERENT METHODS

To display the discriminating ability of dual capsule networks in the proposed method, we conducted the translation between normal images and defect images from six kinds of industrial optical data (Class1-6) in DAGM 2007,

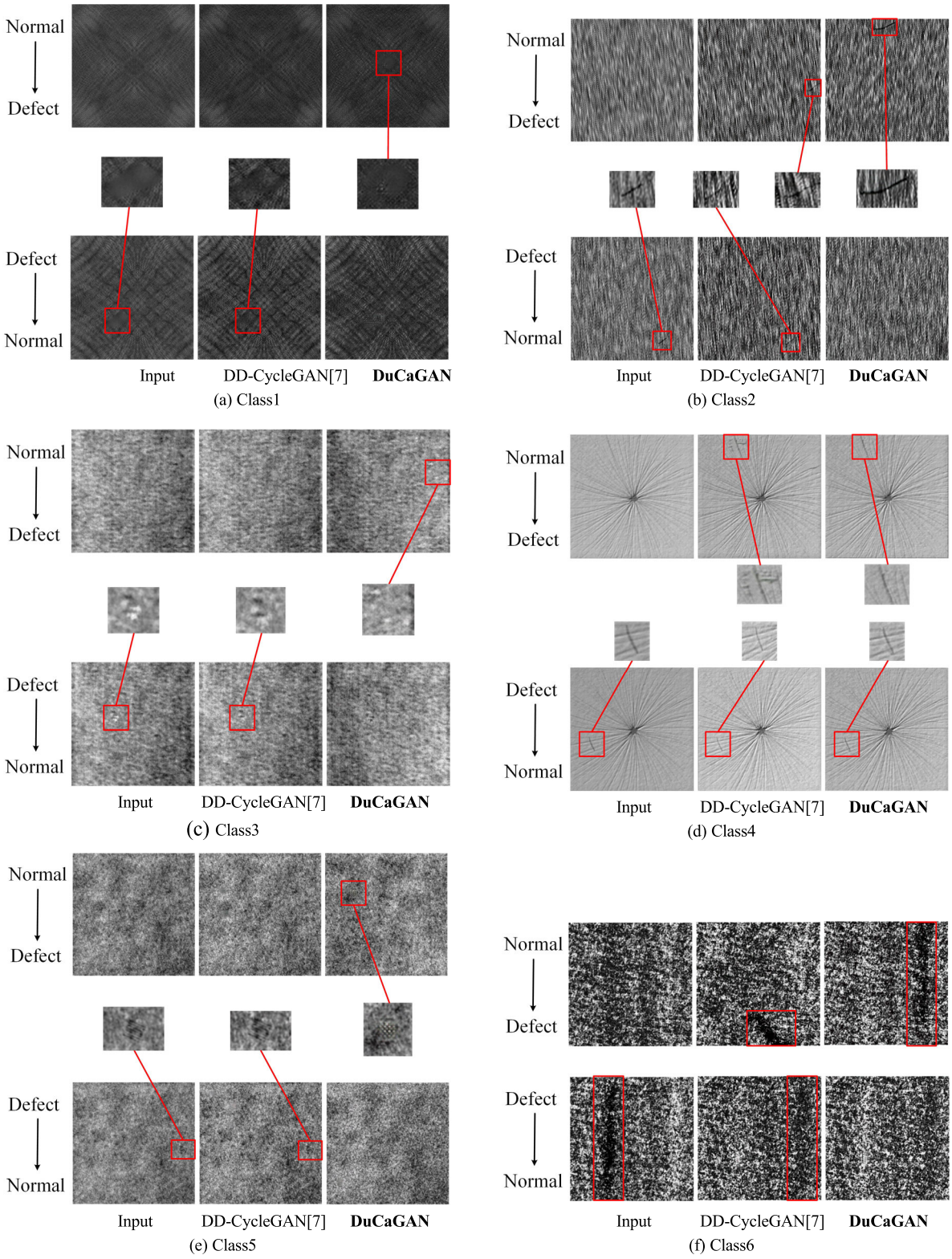

FIGURE 16. Results for normal data ←→ defective data translation from six kinds of industrial optical data (Class1-6) in DAGM 2007.

and compared DuCaGAN with DD-CycleGAN using two general convolutional networks as double discriminators. Fig.16 shows the translation results of different methods. In Figs.16a, 16c, and 16e, one can find that the proposed method is able to generate images with target defects in the normal → defect, but DD-CycleGAN fails to learn the

local defect features in the generated data. Besides, DuCaGAN has a higher ability to remove defective features than DD-CycleGAN in the defect-to-normal translation. As shown in Figs.16 b and 16f, DuCaGAN enables to capture sharper defects than DD-CycleGAN in the normal-to-defective. Moreover, the effect of DD-CycleGAN to remove defects is not as obvious as DuCaGAN in the defect-to-normal task. For images translation of Class4 in DAGM 2007, one can find that all methods are difficult to effectively restore normal images by removing defective features in defect-to- normal translation from Fig.16 d. Although both DD-CycleGAN and DuCaGAN can generate defects in the normal-to-defective translation, there are other unknown features around the defects generated by the DD-CycleGAN in the normal-to-defective. Therefore, DuCaGAN can gener- ate high-quality target images more efficiently than DD-CycleGAN in DAGM 2007. We can conclude that the learning capability of capsule network discriminators in the proposed model outperforms the double discriminator of DD-CycleGAN for the local detailed features.

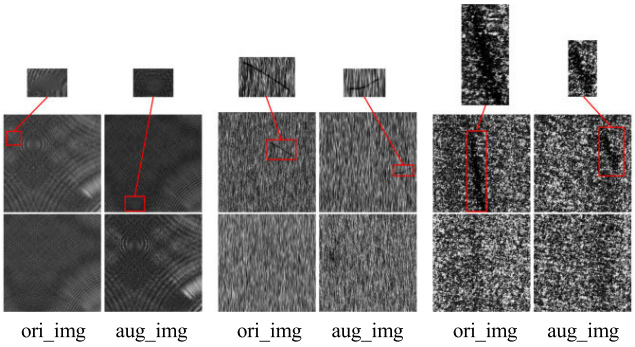

FIGURE 17. The original image (ori_img) and the corresponding image augmented by the proposed DuCaGAN (aug_img) in industrial optical data. The first row is the defective samples, and the second row is the normal samples.

2) ANALYSIS OF IMAGE QUALITY BASED ON DATA AUGMENTATION

To verify that the proposed method can generate reasonable high-quality industrial optical images, we divided the original data into the defective image domain and the normal image domain according to different categories, and then input them into the proposed DuCaGAN to augment the target data. As shown in Fig. 17, the second, fourth, and sixth columns ofthe first row are the defect sample images augmented by the proposed method. One can see from Fig. 17 that the defective image generated by this DuCaGAN contains defect features similar to the original defect image, and the produced entire image has a higher quality. Besides, the second, fourth and sixth columns of the second row are the normal sample images augmented by the proposed method. The global characteristics of augmented normal images are very similar to the original normal samples from industrial optical data. Therefore, the images generated by the DuCaGAN can maintain a higher consistency with the original image from normal samples and defective samples in industrial optical data.

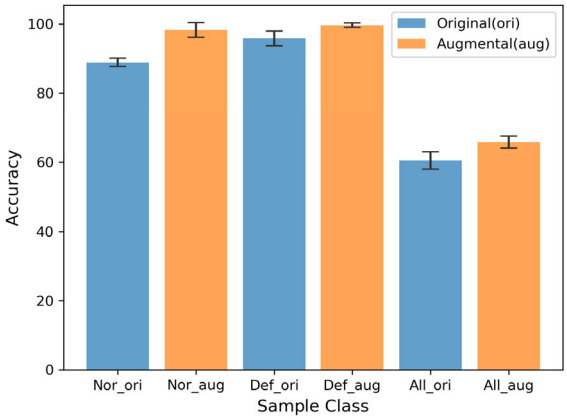

FIGURE 18. Accuracy of industrial optical classification on optical data augmented by the proposed DuCaGAN in DAGM 2007 dataset.

3) APPLICATION OF DATA AUGMENTATION

To illustrate that the images augmented by the DuCaGAN can improve the generalization performance of the image classification model, we trained the image classification model respectively by utilizing the original industrial optical data and the combined data of the original images and the augmented images. We divided the original data and the combined data into different sample groups according to sample types to train the image classification model, as shown in Fig. 18. Nor is the normal images group. Def is the defective images group, and all is the all images group. The statistical classification accuracy of different sample groups after multiple experiments are shown in Fig. 18. For the normal sample group, the classification accuracy of combined data from the proposed method greatly outperforms the original data, but the deviation is higher. This indicates that the quality of generated images is average, although the DuCaGAN can effectively augment industrial optical normal images to improve the generalization performance of the image classification model. For the defect sample group, the improvement of the image classification accuracy is limited, while the deviation is lower, indicating shows that the defective images augmented by the proposed method have higher quality. Finally, for all images sample group including normal images and defective images, the combined data generated by the proposed DuCaGAN not only improve the accuracy of image classification, but also have relatively high image quality according to the small deviation.

Therefore, to a certain extent, the industrial optical data augmented by the DuCaGAN, to a certain extent, improve the over-fitting of image classification model and generalization performance.

IV. CONCLUSION

In summary, we proposed a new unsupervised learning framework by introducing the capsule network and the multi-agent competition mechanism into the generative adversarial networks to solve the image-to-image translation. Two capsule networks were added into the cycle consistent structure as discriminators. Also, to improve the entire full objective and stabilize the training procedure, we modified the previous loss function of the GAN network by combining the margin loss function of the capsule network. Experiments on seven datasets were conducted to compare our proposed DuCaGAN model with the CycleGAN and DCGAN methods. The results reveal that DuCaGAN can produce more realistic images and approximate the true distribution of the target domain more accurately. Furthermore, it not only can preserve the structural information of the source domain, but also can learn the detailed features of the target domain. More importantly, DuCaGAN is applied for the first time to augment the surface defect data from the real industrial field and are obtained than those from other methods. However, our method is time-consuming during training. Although it exhibits the better performance on color and texture features than the previous methods, a further enhancement is still needed in geometrical features.

This paper raises the possibility of studying scientific issues in an unsupervised way. Since it is easy to obtain the unpaired data, a major trend isto solve the problem of image generation and translation in an unsupervised way. In the future, the fusion of multi-domain information and even multi-modal information is an important direction to address the above problems.

REFERENCES

[1] G. Mariani, F. Scheidegger, R. Istrate, C. Bekas, and C. Malossi, "BAGAN: Data augmentation with balancing GAN," 2018, *arXiv:1803.09655*. [Online]. Available: http://arxiv.org/abs/1803.09655

[2] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1125–1134.

[3] M. Johnson-Roberson, C. Barto, R. Mehta, S. N. Sridhar, K. Rosaen, and R. Vasudevan, "Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks?" 2016, *arXiv:1610.01983*. [Online]. Available: http://arxiv.org/abs/1610.01983

[4] Z. Zhao, B. Li, R. Dong, and P. Zhao, "A surface defect detection method based on positive samples," in *Proc. Pacific Rim Int. Conf. Artif. Intell. (PRICAI)*, Aug. 2018, pp. 473–481.

[5] M. Gadermayr, K. Li, M. Müller, D. Truhn, N. Krämer, D. Merhof, and B. Gess, "Domain-specific data augmentation for segmenting MR images of fatty infiltrated human thighs with neural networks," *J. Magn. Reson. Imag.*, vol. 49, no. 6, pp. 1676–1683, Jun. 2019.

[6] A. A. Efros and W. T. Freeman, "Image quilting for texture synthesis and transfer," in *Proc. 28th Annu. Conf. Comput. Graph. Interact. Techn. (SIGGRAPH)*, 2001, pp. 341–346.

[7] J. Zhao, J. Zhang, Z. Li, J.-N. Hwang, Y. Gao, Z. Fang, X. Jiang, and B. Huang, "DD-CycleGAN: Unpaired image dehazing via double-discriminator cycle-consistent generative adversarial network," *Eng. Appl. Artif. Intell.*, vol. 82, pp. 263–271, Jun. 2019.

[8] Y. Shih, S. Paris, F. Durand, and W. T. Freeman, "Data-driven hallucination of different times of day from a single outdoor photo," *ACM Trans. Graph.*, vol. 32, no. 6, p. 200, Nov. 2013.

[9] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, "Diverse image-to-image translation via disentangled representations," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 36–52.

[10] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[11] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*. [Online]. Available: http://arxiv.org/abs/1411.1784

[12] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2223–2232.

[13] S. Ma, J. Fu, C. W. Chen, and T. Mei, "DA-GAN: Instance-level image translation by deep attention generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5657–5666.

[14] A. Gupta, S. Venkatesh, S. Chopra, and C. Ledig, "Generative image translation for data augmentation of bone lesion pathology," 2019, *arXiv:1902.02248*. [Online]. Available: http://arxiv.org/abs/1902.02248

[15] J. Yuan, W. Zhou, and T. Luo, "DMFNet: Deep multi-modal fusion network for RGB-D indoor scene segmentation," *IEEE Access*, vol. 7, pp. 169350–169358, 2019.

[16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[17] Y. Tao, Z. Ling, and I. Patras, "Universal foreground segmentation based on deep feature fusion network for multi-scene videos," *IEEE Access*, vol. 7, pp. 158326–158337, 2019.

[18] J. Song, Y. Guo, L. Gao, X. Li, A. Hanjalic, and H. T. Shen, "From deterministic to generative: Multimodal stochastic RNNs for video captioning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 10, pp. 3047–3058, Oct. 2019.

[19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Dec. 2014, pp. 2672–2680.

[20] J. Song, J. Zhang, L. Gao, X. Liu, and H. T. Shen, "Dual conditional GANs for face aging and rejuvenation," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 899–905.

[21] Y. Li, A. Schwing, K.-C. Wang, and R. Zemel, "Dualing GANs," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Dec. 2017, pp. 5606–5616.

[22] D. He, Y. Xia, T. Qin, L. Wang, N. Yu, T.-Y. Liu, and W.-Y. Ma, "Dual learning for machine translation," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Dec. 2016, pp. 820–828.

[23] T. Kim, M. Cha, H. Kim, J.-K. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Aug. 2017, pp. 1857–1865.

[24] Z. Yi, H. Zhang, P. Tan, and M. Gong, "DualGAN: Unsupervised dual learning for image-to-image translation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2849–2857.

[25] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Dec. 2017, pp. 3856–3866.

[26] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015, *arXiv:1502.03167*. [Online]. Available: http://arxiv.org/abs/1502.03167

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[28] A. Jaiswal, W. AbdAlmageed, Y. Wu, and P. Natarajan, "Capsulegan: Generative adversarial capsule network," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 1–9.

[29] R. Saqur and S. Vivona, "CapsGAN: Using dynamic routing for generative adversarial networks," in *Proc. Sci. Inf. Conf.*, 2019, pp. 511–525.

[30] D. Wang and Q. Liu, "An optimization view on dynamic routing between capsules," presented at the 6th Int. Conf. Learn. Represent. (ICLR), Apr./May 2018.

[31] G.-E. Hinton, S. Sabour, and N. Frosst, "Matrix capsules with EM routing," presented at the 6th Int. Conf. Learn. Represent. (ICLR), Apr./May 2018.

[32] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2016, pp. 3213–3223.

[33] M. Eitz, J. Hays, and M. Alexa, "How do humans sketch objects," *ACM Trans. Graph.*, vol. 31, no. 4, p. 44, 2012.

[34] P.-Y. Laffont, Z. Ren, X. Tao, C. Qian, and J. Hays, "Transient attributes for high-level understanding and editing of outdoor scenes," *ACM Trans. Graph.*, vol. 33, no. 4, p. 149, 2014.
[35] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*. [Online]. Available: http://arxiv.org/abs/1511.06434
[36] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980
[37] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Dec. 2016, pp. 2234–2242.
[38] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8789–8797.

GUIFANG SHAO received the B.S., M.S., and Ph.D. degrees from Chongqing University, China, in 2000, 2003, and 2007, respectively, all in control theory and control engineering.

She is currently an Associate Professor with the Department of Automation, Xiamen University, China. Her research interests include robot control, DNA microarray image processing, and pattern recognition.

MENG HUANG received the B.S. degree in mechanical engineering and automation from the Changshu Institute of Technology (CIT), China, in 2015. He is currently pursuing the M.S. degree in pattern recognition and intelligent systems with the Department of Automation, Xiamen University, China.

His research interests include generative adversarial networks, deep learning, and computer vision, especially in image translation.

FENGQIANG GAO received the B.S. and M.S. degrees in detection technology and automatic equipment from Xiamen University, China, in 2007 and 2010, respectively, where he is currently pursuing the Ph.D. degree in systems engineering with the Department of Automation.

His research interests include deep learning and computer vision detection.

TUNDONG LIU received the M.S. degree in control theory and control engineering from the Gansu University of Technology, China, in 2000, and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, China, in 2003.

He is currently a Professor with the Department of Automation, Xiamen University, China. His research interests include robot control and vision detection.

LIDUAN LI received the B.S. degree in automation from the Ocean University of China (OUC), China, in 2017. She is currently pursuing the M.S. degree in systems engineering from the Department of Automation, Xiamen University, China.

Her research interests include reinforcement learning, computer vision, and robot control.