IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 7, 2020, accepted July 23, 2020, date of publication July 27, 2020, date of current version August 7, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3012185

VSA-CGAN: An Intelligent Generation Model for Deep Learning Sample Database Construction

PENG ZHANG[1], XIA HUA[1], XINQING WANG[1], TING RUI[1], (Senior Member, IEEE), HAITAO ZHANG[1], FAMING SHAO[1], AND DONG WANG[1,2]

[1]College of Field Engineering, Army Engineering University of PLA, Nanjing 210007, China
[2]Second Institute of Engineering Research and Design, Kunming 650222, China

Corresponding author: Xia Hua (huaxia120888@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61671470, in part by the National Key Research and Development Program of China under Grant 2016YFC0802904, and in part by the Postdoctoral Science Foundation Funded Project of China under Grant 2017M623423.

ABSTRACT In order to solve the problem of model accuracy reduction caused by the difficulty of obtaining specific training samples or the insufficient number of samples in the application of existing object detection and recognition model based on deep learning, this article proposes a conditional generative adversarial network model (VSA-CGAN), which integrates the self-attention mechanism of visual perception to optimize the inference of object attention feature maps, so as to learn the global information of the image and the detailed features of the object. It is designed to add conditional features in the generator and the discriminator, associate the specific dimensions of the data with the semantic features, and explicitly indicate the model to generate the corresponding object signature category information, so as to generate the feature representation of the image which is more suitable for the distribution of the original data. The model in this article has completed numerical experiments on several general standard data sets, and compared with several mainstream generative adversarial network models in image data augmentation performance. The experimental results show that the generation model in this article has excellent object simulation ability and strong application prospects.

INDEX TERMS Generative adversarial network, attention mechanism, visual salience, object simulation, deep learning, data augmentation.

I. INTRODUCTION

Generative adversarial network (GAN) is an optimized generation model proposed by Goodfellow in 2014 based on the idea of antagonistic competition. It is developed on the basis of deep generation model, but is highly different from previous models.

The GAN consists of a generator network G and a discriminator network D. The goal of the generator is to fit the sample data and generate the simulation data, while the goal of the discriminator is to distinguish the true and false data. The network structure of the generator and the discriminator is multi-layer perceptron.

Given the real sample $\{x^1, \cdots, x^n\}$, the data of which is published as p_x, and the noise set $\{z^1, \cdots, z^m\}$ is obtained by random sampling from another predefined distribution p_z. Set the input of generator to z and the output of generator to $G(z)$. Input the generated data $G(z)$ and the real data x into the discriminator (the scale can be adjusted according to the experimental situation). The output of the discriminator is a one-dimensional scalar, which represents the probability that the input is true. According to the different input, it is represented as $D(x)$ and $D(G(z))$ respectively. In the ideal case of $D(x) = 1$ and $D(G(z)) = 0$, the network optimization process can be described as a minimax game problem about the value function $V(D, G)$ [1], and the objective function is shown in Formula (1):

$$\min_G \max_D V(D, G) = \mathrm{E}_{x\sim p_x}\left[\log D(x)\right] + \mathrm{E}_{z\sim p_z}\left[\log\left(1 - D(G(z))\right)\right] \quad (1)$$

If the data release of the generated data $G(z)$ is expressed as p_G, then the "binary minimax" problem has a global optimal solution, that is $p_G = p_x$. The training process of generator and discriminator is alternating. When one side's parameters are updated, the other side's parameters are fixed.

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

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

The main difference between CGAN [2] and GAN is that CGAN improves unsupervised generative adversarial network to supervised generation countermeasure network model, introduces conditional variable c to both generation network and discrimination network, adds additional information to guide the generation process of data, and generates samples of specified categories. The objective function of CGAN is shown in formula (2).

$$\min_G \max_D V(D, G) = E_{x\sim p_x}\left[\log D(x|c)\right] + E_{z\sim p_z}\left[\log(1 - D(G(z|c)))\right] \quad (2)$$

The conditional variable c is the added supervision information. CGAN receives conditional variable c and random noise z as input. Generally, c can be any other auxiliary information, such as type label and other data types, and c is input to the discriminator and generator as additional input layer to perform adjustment.

The problem GAN has to solve is the way to train samples and thus learn the probability distribution principles of data. In the processing of learning algorithm, GAN draws inspiration from the zero-sum game theory (under strict competition, one party's income inevitably means the other party's loss, the sum of the profits and losses of each party in the game is always zero, and there is no cooperation between the two parties), and so its network consists of a generative model and a discriminant model [1], [2]. The network integrates the generative model with the discriminant model skillfully, and the prior one models the input parameters and generates data while the latter one identifies the authenticity of the data, thus making it possible for the two models to compete and promote each other. Compared with the commonly used in-depth learning model of Variational Autoencoders, GAN does not require the presupposition that the data obeys a prior distribution, so it has significant advantages. GANS network is in the stage of imagination creation, and because of unique data generation ability, it has become a hot topic for contemporary researches learning. Therefore, a hot topic in this field has traditionally been how to improve the quality of the generated image in AI researches [3], [4].

Based on the latest research progress of GAN, this article proposes a new generative adversarial network—Conditional Generative Adversarial Network based on self-attention mechanism and visual perception (VSA-CGAN), for image data simulation application. The main work can be summarized as follows:

1) Firstly, a cross-correlation self-attention module is designed to balance the computational efficiency and statistical efficiency of the network as well as the ability to simulate remote dependence, to learn the global information and objective characteristics of the image. The self-attention feature map is reasoned and optimized by convolution's long-term and short-term memory network, thus making the model focus more on objects' structure and detailed features, and improving the generating ability of the model.

2) The similarity of conditional feature learning distribution error is introduced into the generator network and the discriminator network as the supervision information, which makes the feature representation of the specific type of image generated by the generation model more robust than the original data distribution.

3) The L_1 loss function is introduced to measure the pixel level difference loss of the image, which makes the network pay attention to the feature information of the image as well as the reconstruction of the image pixel information. The introduction of L_1 loss function also makes the network obtain better performance and convergence speed.

The VSA-CGAN model has completed numerical experiments on multiple different data sets, and has been compared with several currently mainstream generation adversarial network models in image data simulation performance. The experimental results show that the generated model is highly applicable since it has excellent simulation ability, and its generated image data can effectively realize the purpose of data augmentation while dealing with small samples, specific objects, and so on.

II. RELATED WORK

DCGAN [5] is a typical improvement in the early development of GAN. Convolutional neural network (CNN) is a commonly used network structure in image processing objects, which is considered to be able to automatically extract image features [6]. DCGAN replaces the full connection layer in the generator with deconvolution layer, achieving good performance in image generation objects. Therefore, nowadays, when using GAN for image generation objects, the default network structure is generally similar to DCGAN settings.

Self-Attention Generative Adversarial Networks (SA-GAN) can use clues from all feature locations to generate detailed information (traditional convolutional genetic algorithm only generates high-resolution detailed information based on spatial local points in low-resolution feature mapping). The spectral normalization is applied to the GAN generator to improve the training dynamics. The Inception Score (IS) of the image generated by 128 × 128 ImageNet can reach 52 points [7].

Big-Gan, developed by DeepMind, introduces the idea of orthogonal regularization into Gan. By increasing the number of parameters (increasing channels) by 2-4 times and expanding the batch size by 8 times, Gan can get the maximum performance improvement. By using truncation techniques, training can be more stable, but it needs to balance diversity and fidelity. By existing and other novel technologies, the combination of techniques can ensure the stability of training, but the accuracy will also decline, so it needs to balance the performance and training stability. In the Imagenet dataset, the perception score is more than 100 points higher than the SA-GAN model [8].

On the premise of ensuring the quality of the generated images, Feng Yong and Zhang Chun follow the continuous updating and iteration in drawing to improve the diversity

of the generated samples and enhance the semantics of the samples. At the same time, Wasserstein distance is introduced, and a Wasserstein image cyclic Generative adversarial network model, abbreviated as WIRGAN (Wasserstein Image Recurrent Generative Adversarial Network Model) [9] is proposed.

Ji and Ma [10] proposed an image generation method of conditional self-attention generative adversarial network. This network combines the advantages of self-attention generation network, adds additional conditional features to the generator and discriminator, explicitly indicates that the model generates the corresponding symbolic category information, associates the specific dimensions of the data with the semantic features, and extracts the generation model by this method.

Jie Feng and Xueliang Feng [11] proposed a symmetric convolutional GAN based on collaborative learning and attention mechanism (CA-GAN). In CA-GAN, the generator and the discriminator not only compete but also collaborate. The shallow to deep features of real multiclass samples in the discriminator assist the sample generation in the generator. In the generator, a joint spatial–spectral hard attention module is devised by defining a dynamic activation function based on a multi-branch convolutional network. It impels the distribution of generated samples to approximate the distribution of real HSIs both in spectral and spatial dimensions, and it discards misleading and confounding information. In the discriminator, a convolutional LSTM layer is merged to extract spatial contextual features and capture long-term spectral dependencies simultaneously. Finally, the classification performance of the discriminator is improved by enforcing competitive and collaborative learning between the discriminator and generator.

Chen *et al.* [12] proposed a coarse-and-fine structure, which can extract coarse features with a larger receptive field to guarantee the accuracy of global semantic information, and can simultaneously extract fine features with a smaller receptive field at multiple levels to serve as a supplement in a parallel manner. To avoid artifacts caused by the missing part of image, they flexibly use the gating mechanism and propose an interleaved gated residual block (IGRB) to encourage the useful information flow through the neural network. Moreover, they proposed a channel and spatial attention block (CSAB) to alleviate the influence of redundant information and better model the long-range dependency between different regions in the image. Extensive experiments on faces, natural objects and scenes demonstrate that their method outperforms the existing state-of-the-art methods.

III. THEORETICAL METHOD

A. STRENGTHENING OBJECT SELF-ATTENTION MECHANISM

The Attention Mechanism stems from the study of human vision. In order to make rational use of limited visual information processing resources, humans need to select a specific part of the visual area and then focus on it. The attention mechanism has two main aspects: deciding which part of the input to focus on; assigning limited information processing resources to the important part [13]–[17].

SAGAN brings attention mechanism into the image generation task of GAN. Self-attention mechanism shows a better balance among the ability to simulate remote dependence, computational efficiency and statistical efficiency. The self-attention module takes the weighted sum of features at all locations as the response of the location, where the weight or attention vector is only calculated at a small calculation cost. Although SAGAN can learn the distribution rule of the overall geometric features of the image based on the self-attention module to a certain extent, the self-attention model of SAGAN is not precise enough to learn the distribution of the structural information and geometric features of the object itself, resulting in the poor effect of generating the detailed features of the object in the generated image, the deviation of the geometric distribution between the key structures of the object, and SAGAN still adopts an unsupervised learning method, and the designed attention model has a high demand for the number of training data, which greatly limits the performance and application prospects of the network [7].

On the basis of self-attention model, we design a new kind of strengthening object self-attention network (SOSA -Net). By introducing conv-LSTM network, we optimize the self-attention feature map, which makes the model more perfect for the learning of the structural features and detailed features of the object itself. The body structure is shown in Figure 1.

After training, the network model gathers the information of high-level self-attention feature map and the rich spatial and detailed features of the underlying network through the feature Pyramid [19] strategy, and focuses on and continuously refines the object in the complex background. As shown in the figure, the model is calculated in a top-down way, integrating information from the early layers in turn. Multiple conv-LSTM networks [20] (green blocks in the figure) are stacked to construct more meaningful feature expression results with circular connection. We use the order and memory characteristics of LSTM to process features in an iterative approach. At a certain level, conv- LSTM abandons the feature of small amount of information and strengthens the feature of large amount of information, thus gradually improving the feature map of reasoning object enhanced self-attention.

$f(x)$, $g(x)$ and $h(x)$ are all common 1×1 convolutions, and their difference lies only in the size of the output channel: the output is transposed and multiplied by the output, then an attention map is obtained by normalizing the soft Max function; SAGAN multiplies this attention map and the convolution result by pixels [7] to get the adaptive attention maps.

SAGAN used the proposed attention model in the generator and discriminator, and finally used alternate training to minimize adversarial loss.

$$L_{DS} = -\mathrm{E}_{(x,y)\sim p_{data}}[\min(0, -1 + D(x, y))] - \mathrm{E}_{z\sim p_z, y\sim p_{data}}[\min(0, -1 - D(G(z), y))] \quad (3)$$

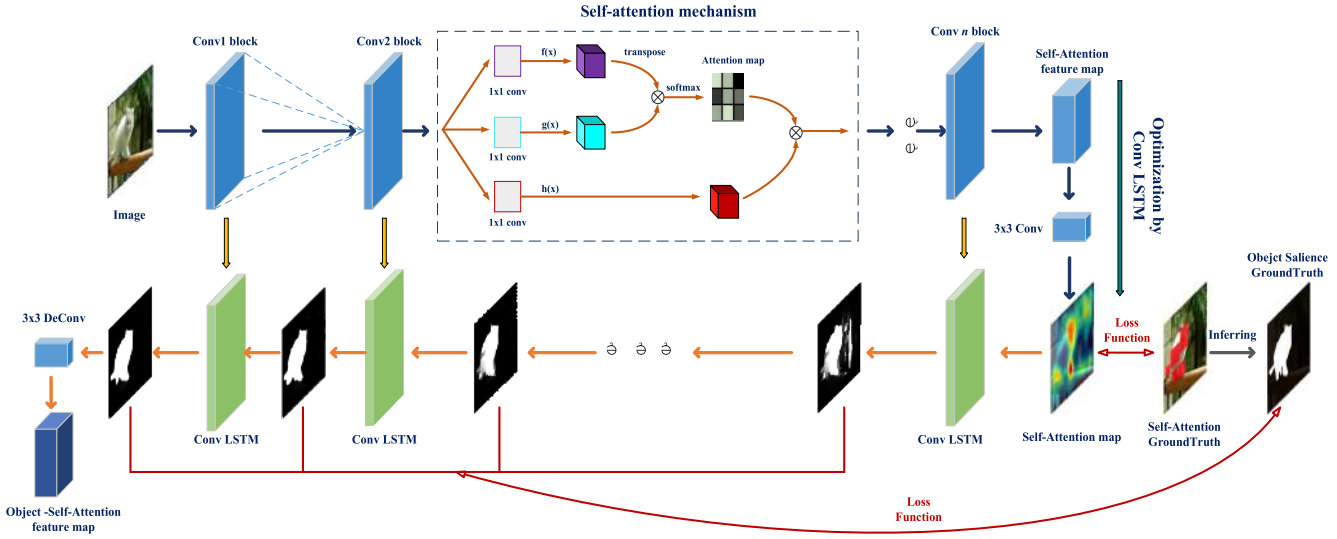


FIGURE 1. SOSA-Net overall network structure and data processing flow. This model does computation from top to bottom, integrating information from the early layers one by one. Multiple conv-LSTM networks (green blocks marked) are stacked together to construct more meaningful features to express the results. We use the order and memory characteristics of LSTM to process features iteratively. At a certain level, conv-LSTM abandons the feature of small amount of information and strengthens the feature of large amount of information, thus generating a stepwise improved and enhanced self-attention feature map.

$$L_{GS} = -\mathrm{E}_{z\sim p_z, y\sim p_{data}} D\left(G\left(z\right), y\right) \tag{4}$$

However, SAGAN's multi-feature fusion rule ignores the cross-correlation characteristics of the three feature spaces f, g, h of the image, that is, the cross-correlation characteristics of the three weight matrices W_f, W_g, W_h. For different task scenes and object types, the influence degree of the three feature spaces on the final Attention Map generation effect is different. The higher the matching degree of the three feature spaces, the better the final overall generated attention map effect is. Therefore, in the training process of the three weight matrices, we introduced a cross-correlation mechanism to ensure that the training results have good matching degree. Cross-correlation function is a concept in signal analysis, which indicates the degree of correlation between two time series, that is, describes the degree of correlation between the values of two different signals at any two different times. The definitions of cross-correlation on continuous function and discrete function are equation (5) and equation (6) respectively.

$$\left(\dot{f} * \dot{g}\right)(\psi) = \int_{-\infty}^{+\infty} \dot{f}^*(t)\, \dot{g}(t+\psi)\, \mathrm{d}t \tag{5}$$

$$\left(\dot{f} * \dot{g}\right)(n) = \sum_{-\infty}^{+\infty} \dot{f}^*[m]\, \dot{g}[m+n] \tag{6}$$

where $\dot{f} * \dot{g}$ represents the cross-correlation function of $\dot{f}$ and $\dot{g}$ and a $\dot{f}^*$ represents the complex conjugate function of $\dot{f}$. Cross-correlation function is similar to convolution operation and is also the sliding multiplication of two sequences, which reflects the degree to which the two functions match each other in different relative positions [18]. Because the training process of each weight matrix belongs to discrete function, equation (7) is used to calculate and the results which are normalized by discrete standardization.

$$x^* = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{7}$$

The cross-correlation values between two of the three weight matrices are expressed in terms of $R\left(x_f, x_g\right)$, $R\left(x_f, x_h\right)$, $R\left(x_g, x_h\right)$. After the weight gets loss function converges, the average cross-correlation value is used to select the optimal weight matrix, that is, equation (8):

$$R = \left(R\left(x_f, x_g\right) + R\left(x_f, x_h\right) + R\left(x_g, x_h\right)\right)/3 \tag{8}$$

Conv-LSTM extends the traditional fully connected LSTM to handle spatial features. Basically, this is achieved by using convolution instead of dot product in LSTM equation. Conv-LSTM has a convolution structure in the input to the state and the state-to-state transition, which can preserve the spatial information of convolution feature map, thus enabling our network to generate pixel-level labels. Similar to the traditional gate LSTM, conv-LSTM uses memory cells and gates to control information flow. It works by sequentially updating the internal state H and memory cell C through the values i, f', c of three sigmoid gates. In the t step, when the input X_t arrives, if the input gate i_t is activated, the included information of X_t will be accumulated in the memory cell, and if the forgotten gate f_t' is turned on, the state C_{t-1} of the previous memory cell will be forgotten. Whether the latest cell state C_t should propagate to the final state H_t is further controlled by the output gate o_t [20]. Here, we use the recursive nature of LSTM to iteratively optimize the salient features of static images, instead of using LSTM to model the time dependence of sequence data.

B. STRENGTHENING OBJECT SELF-ATTENTION GAN MODEL

We combine the features of self-attention prior graph P_s and convolution layer as the input of conv-LSTM. In each time step, conv-LSTM is trained, and salient objects are inferred by using fixed information knowledge, and the features are sequentially optimized according to the updated storage unit and hidden state. Therefore, the features are reorganized to better represent the significance of the object. First, we compress the characteristic response of the convolution layer through convolution layer of multiple filters to reduce the calculation cost, and use sigmoid activation to regularize the characteristic response so that it is within the same range of P_s. Then, the self-attention prior graph P_s is connected with the compressed features and input to conv-LSTM. We apply the 1×1 and 3×3 combined convolution kernels to the final conv-LSTM output H to obtain the inference object enhanced self-attention feature map Q.

In order to evaluate the significance model, several different measurement standards have been proposed. We adopt the real prominent object annotation S proposed in the document [20], and thus we can obtain the conv-LSTM total loss function defined as equation (9):

$$L_{Sal}(S,Q) = L_C(S,Q) + \alpha_1 L_P(S,Q) + \alpha_2 L_R(S,Q) + \alpha_3 L_F(S,Q) + \alpha_4 L_{MAE}(S,Q) \quad (9)$$

where the balance parameters are set to $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0.1$, L_C is the weighted cross entropy loss function, which is the main loss function of conv-LSTM model as equation (10):

$$L_C(S,Q) = \frac{1}{N}\sum_x \begin{pmatrix} \vartheta \cdot (1-s_x) \cdot \log(1-q_x) \\ +(1-\vartheta) \cdot s_x \cdot \log q_x \end{pmatrix} \quad (10)$$

where N represents the total number of image pixels, and $s_k \in S$, $q_k \in Q$. ϑ refers to the ratio of S significant pixels in the real value, and the weighted cross entropy loss function handles the imbalance between prominent and non-prominent pixels. L_P, L_R, L_F are used to calculate the similarity of precision, recall and F-measure scores:

$$L_P(S,Q) = -\sum_x s_x \cdot q_x \Big/ \left(\sum_x q_x + \varepsilon\right) \quad (11)$$

$$L_R(S,Q) = -\sum_x s_x \cdot q_x \Big/ \left(\sum_x s_x + \varepsilon\right) \quad (12)$$

$$L_F(S,Q) = -\frac{\left(1+\beta^2\right) \cdot L_P(S,Q) \cdot L_R(S,Q)}{\beta^2 \cdot L_P(S,Q) + L_R(S,Q) + \varepsilon} \quad (13)$$

where $\beta^2 = 0.3$ is the setting according to document [19], ε is a regularization constant. Because precision, recall and F-measure are similarity measures, higher values are better, so negative values are used to minimize. L_{MAE} is derived from the mean absolute error (MAE) metric, which calculates the difference between the significance graph Q and the truth graph S.

$$L_{MAE}(S,Q) = \frac{1}{N}\sum_x |s_x - q_x| \quad (14)$$

After obtaining the object saliency map Q inferred from the attention prior map P, we unsampled Q and fed it to the next conv-LSTM to obtain the compression feature from the conv n-1 layer for more detailed optimization. The above process iterates layer by layer to the conv 1 layer. in short, the model can effectively infer the salient features of the learning object, which is due to 1) the learnable self-attention mechanism, 2) iteratively updating the salient features and the cyclic architecture, and 3) effectively merging the spatial rich information from the lower layers in a top-down manner.

We set $y_k^A \in \{0,1\}$ and $y_k^S \in \{0,1\}$, and point out whether we have attention annotations G_k and object saliency masks S_k for the training image sequenced by k. Finally, the loss function of the whole generation network and detection network is shown in equation (15) and equation (16):

$$L_G = \sum_{k=1}^{K} y_k^A \cdot L_{GS} + \sum_{k=1}^{K} y_k^S \cdot \sum_{v=1}^{n} L_{Sal}\left(S_k^v, Q_k^v\right) \quad (15)$$

$$L_D = \sum_{k=1}^{K} y_k^A \cdot L_{DS} + \sum_{k=1}^{K} y_k^S \cdot \sum_{v=1}^{n} L_{Sal}\left(S_k^v, Q_k^v\right) \quad (16)$$

The lack of ground authenticity in the corresponding tasks was corrected by using $y_k^A \in \{0,1\}$ and $y_k^S \in \{0,1\}$ as indicators. That is, when no annotations are provided, the error does not propagate back. Here, v represents the layer which is in conv-LSTM. Through the hierarchical loss function, each layer in the model can directly access the gradient of the loss function, so as to achieve implicit in-depth monitoring.

C. GENERATION NETWORK AND DISCRIMINATION NETWORK OF VSA-CGAN MODEL

Based on the supervision idea of CGAN, conditional variables are introduced into generators and discriminators, and additional information is used to guide the data generation process. Combined with the advantages of SOSA-Net model, the structure of VSA-CGAN is shown as Figure 2:

The optimization function of VSA-CGAN changes the prior probabilities x and y in L_{DS} and L_{GS} into the posterior probabilities $x\,|c$ and $y\,|c$ as equation (17) and equation (18):

$$L_{DS} = -\mathrm{E}_{(x,y)\sim p_{data}}\left[\min\left(0, -1 + D\left(x\,|c,\ y\,|c\right)\right)\right] - \mathrm{E}_{z\sim p_z, y\sim p_{data}}\left[\min\left(0, -1 - D\left(G\left(z\,|c\right), y\,|c\right)\right)\right] + L_1 \quad (17)$$

$$L_{GS} = -\mathrm{E}_{z\sim p_z, y\sim p_{data}} D\left(G\left(z\,|c\right), y\,|c\right) \quad (18)$$

L_1 is the regularization loss function [21] used to measure the pixel level difference loss of the image, which makes the image generated by the generator smoother and faster convergence. Expressed as equation (19), where λ is the regularization coefficient.

$$L_1 = \lambda \sum_j |\omega_j| \quad (19)$$

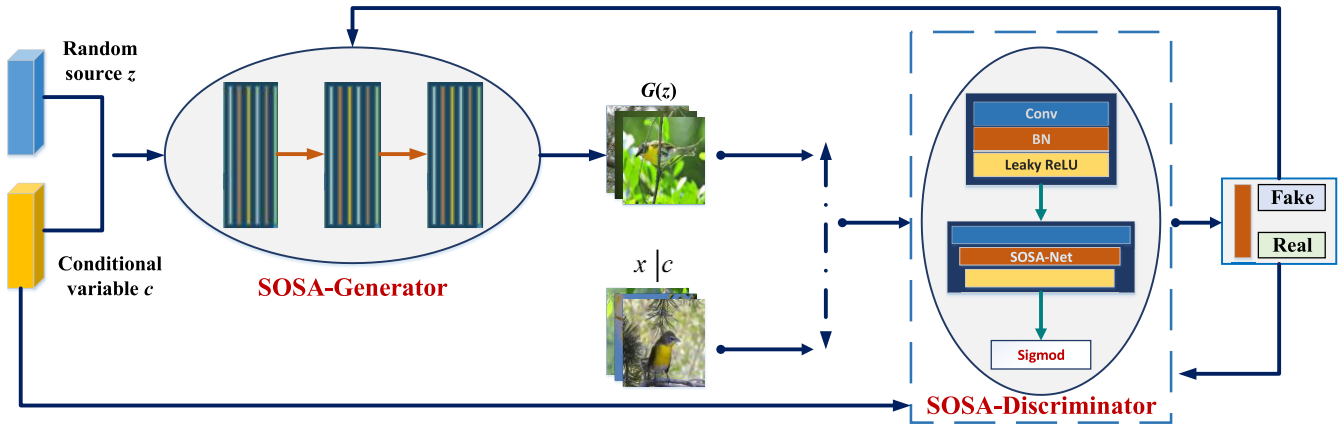


FIGURE 2. Network structure of VSA-CGAN. The generator of the model receives random noise and conditional feature as input, generates simulated picture data, and then inputs the generated picture and original conditional feature into the discriminator, and the discriminator receives the real picture with conditional feature as input.

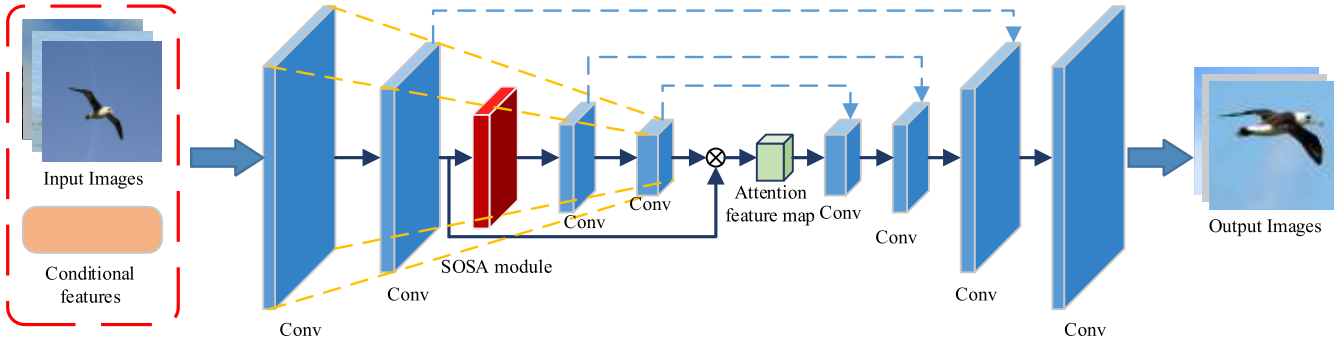


FIGURE 3. VSA-CGAN generator network structure diagram.

Figure 2 also shows the confrontation training process of the model, which fixes the discriminant model, sends the generated image data and constraints into the discriminant model at the same time, and guides the generated model to optimize according to the discriminant results. Fixed generative model, the generated image data and constraints are made available for the discriminant model at the same time, which makes the discriminant model more sensitive to distinguish the difference between the generated image data and the supervision label. In this way, the training is completed until the trained discriminant model cannot distinguish the true and false of the generated images.

The VSA-CGAN generator uses the U-Net network [22]. The structure is shown in Figure 3. The generation network mainly has a convolution feature extraction module, the deep enhancement object self-attention residual module SOSA-Net, and the image up sampling module. The image reconstruction module consists of four parts. U-Net is a full convolutional structure neural network for image generation. The difference between the U-Net and the Encoder-Decoder network is that U-Net establishes a connection for the corresponding feature map before and after decoding. The normal encoder is down sampled to a lower dimension and then up sampled to the original size, while U-Net splices the corresponding feature maps in the encoding phase and the decoding phase through the channel, thus saving pixel levels at different resolutions. The details of the content make the images generated by the generative model and the supervised tags more similar in detail.

The discriminator model uses the Patch-GAN classifier [23]. In the traditional GAN, the discriminant model maps the image to be discriminated to a vector to determine the generation effect of the generative model. When Patch-GAN is used for CGAN, the image to be discriminated (image generated by the generator) and the supervised label are divided into a plurality of area blocks of size $N \times N$. Different blocks are independent of each other, and only one corresponding block needs to be input at a time, and the generation effect of the block is obtained by a convolution operation. The output results of all the blocks are combined to obtain a trueness feature map, and finally the average of the feature map is considered as the output of the final discriminant model. This allows the dimensions of the input mage to be reduced, the time required for calculations and the number of parameters to be decreased, and images of any size to be calculated. The structure of the discriminator model is shown in Figure 4.

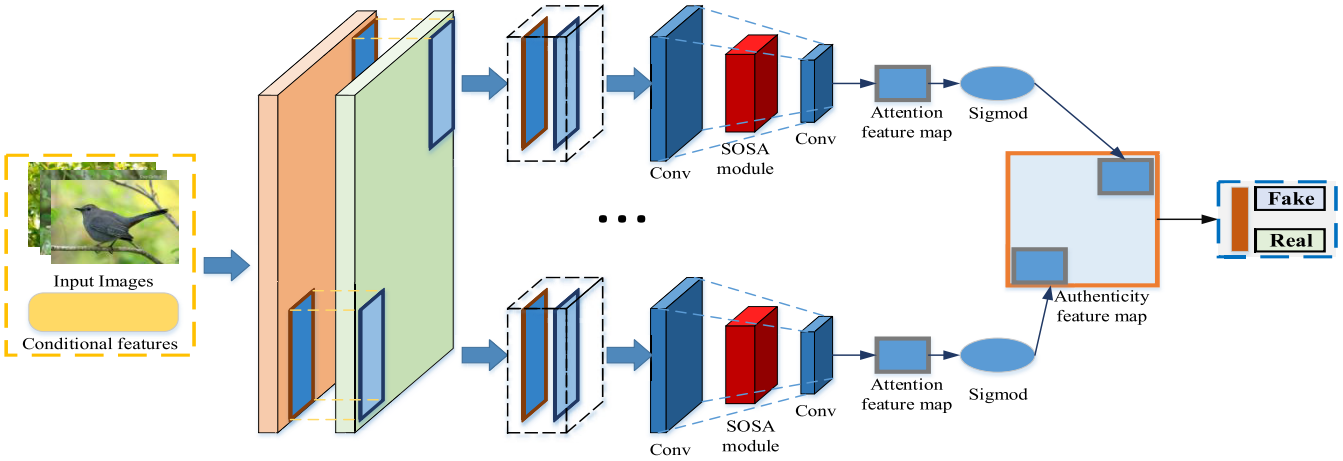


FIGURE 4. VSA-CGAN discriminator network structure diagram.

IV. EXPERIMENTS

In this chapter, we implement ablation research to analyze and verify the impact of different strategies. At the same time, in order to verify the overall advancement of this model, we compared the advanced models based on deep learning in recent years.

A. EXPERIMENTAL DATA SET

In order to fully and concretely reflect the performance and characteristics of the model, we selected different types of standard data sets, including standard handwritten recognition database MNIST [24], face data set CelebA [25] and four popular object detection datasets HKU-IS [26], KITTI [27], PASCAL-S [28], MIT 1003 [29].

In addition to these data sets, in order to verify the performance of the model in some special task scenarios in practical application, we take the simulation task of military object image as an example, and collect and screen a total of 3000 images on the Internet through Google search engine to form the MOD (Military Object image Data set). It falls into three categories: ZTZ-96 main battle tank, M1A2 main battle tank and AH-64-armed helicopter. Each image is classified and marked with attention. These images are normalized into three different size datasets MOD1, MOD2 and MOD3 with 64×64, 128×128, 256×256 size separately.

In this article, the test platform CPU is Intel Core i7 6700, the deep learning operating environment is TensorFlow 2.0, cuDNN 7.4.1, CUDA 10.0, Python version 3.7, graphics card is NVIDIA Titan X, and Ubuntu x64 operating system is used.

B. EXPERIMENTAL VALIDATION OF SOSA-NET

In order to explore the effectiveness of the proposed SOSA-Net mechanism, we use the SN-GAN [30] model as the basic model by adding SOSA modules at different stages of the SN-GAN generator and discriminator. Several GAN models were constructed and compared with the original common SN-GAN model and SA-GAN with self-attention mechanism. At the same time, in the improved model, we use the residual blocks with the same number of parameters to replace the SOSA-Net module in an equal-position, which further validates the effectiveness of SOSA-Net on the improvement of the GAN model.

Regarding the evaluation criteria, we choose two indicators, IS [31] (inception score) and FID [32] (Fréchet Inception Distance) to evaluate different GAN models. In our experiment, 50K samples were randomly generated for each model to calculate the initial score and FID. By default, spectral normalization is used for layers in generators and discriminators, using conditional batch standardization in generators and projection in discriminators. For all models, we use Adam optimizer [33]. By default, the learning rate of discriminator is 0.0004, the learning rate of the generator is 0.0001, and each batch is 64 samples.

In order to explore the effect of the proposed self-attention fusion mechanism SOSA-Net, we construct several GAN models by adding SOSA-Net at different stages of generator and discriminator, and compare the performance with the original SN-GAN model and SA-GAN with self-attention mechanism. What's more, we improve the model by adding SOSA-Net, and the module is replaced by residual blocks with the same number of parameters, which further verifies the effectiveness of SOSA-Net in improving the effect of GAN model. As shown in Table 1, we show the FID scores (column 2), IS scores (column 3) and the best FID scores (column 4), as well as the maximum IS scores and corresponding FID scores (score at the max IS, column 5) for effect evaluation using different models of the ImageNet dataset in generating 256 *256 resolution images, so that we can clearly see how much the diversity will suffer if the maximization of quality is to be achieved [8].

In the Table 1, there is the improved GAN model with SOSA-Net mechanism. Compared with the SN-GAN model without SOSA-Net mechanism, the IS score of the improved GAN model increases by 43.13, that is, from 39.23 to 82.36. Meanwhile, the PID score decreases by 11.62, that is, from 25.73 to 14.11. The overall performance of the model is

TABLE 1. Comparison of SOSA-Net on GANs.

Model		FID	IS	(min FID)/IS	FID/ (max IS)
SN-GAN		25.73	39.23	N/A	N/A
SA-GAN		19.86	55.68	N/A	N/A
SOSA-Net	*feat* 8	21.32	40.63	13.2±0.5/103.4±1	46.2±5/143.4±1
	feat 16	20.81	42.59	11.6±0.2/119.2±1	40.5±4/161.5±3
	feat 32	19.23	53.97	9.8±0.4/131.3±4	29.1±1/178.7±1
	feat 64	13.18	62.32	8.4±0.1/148.3±3	27.2±6/183.2±0
	feat 128	14.11	82.36	8.6±0.1/168.3±2	28.1±8/191.6±5
Residual	*feat* 8	44.61	28.84	42.2±0.2/65.2±2	89.6±8/71.3±4
	feat 16	29.16	31.43	31.8±0.6/73.1±1	76.2±6/81.6±1
	feat 32	25.21	40.03	24.2±0.4/75.4±0	65.1±3/89.1±2
	feat 64	21.96	51.92	21.7±0.3/83.4±3	55.8±7/91.5±5
	feat 128	23.89	45.91	22.3±0.2/78.5±2	59.6±5/96.3±3

improved significantly, which verifies the effectiveness of the SOSA-Net mechanism for GAN model optimization. Compared with the SA-GAN model, the IS score increased from 55.68 to 82.36, and the PID score decreased by 5.75, from 19.86 to 14.11. The overall performance of the model was significantly improved, which verified the effectiveness of using conv-LSTM in SOSA-Net model to optimize the reasoning strategy of self-attention feature map.

As shown in the Table 1, GAN models with SOSA-Net mechanism in feat 64 and 128 show better performance than the one in feat 8 and 16. For example, compared with the GAN model with feat 16, the IS score of the one with feat 128 increased from 42.59 to 82.36, FID score decreased from 20.81 to 14.11, So the overall performance of the model improved significantly: the optimal FID score decreased by about 3.0 and its corresponding IS score (column 4) increased by about 49.1; the best IS score increased by about 30.0 and the corresponding FID score (column 4) decreased by about 12.4.

As shown in Table 1, the SOSA-Net module can also achieve better results than the residual block with the same number of parameters. For example, when we replaced the self-paying block with a residual block in feat 8, training instability occurred, which led to a significant reduction in network performance (for example, the FID increased from 21.32 to 44.61). Even for the case where the training is going smoothly, replacing the self-paying block with the residual block still results in a worse result in terms of FID and IS score. (For example, the FID in the feat 64 is increased from 13.18 to 21.96). This comparison demonstrates that the performance improvements provided by the SOSA-Net strategy are not only due to the increased depth and capacity of the model.

In order to further evaluate the effectiveness of SOSA-Net in visual attention, we first verify the performance of SOSA-Net for FP (fixation prediction) tasks. The purpose of this experiment is to study the validity of fixed map in advance learning, rather than to compare it with the most advanced FP model. Then we evaluate the performance of SOSA-Net in primary SOD (Salient object detection) tasks. For FP tasks, we use five typical measures: Normalized Scan path Saliency (NSS), Similarity Metric (SIM), Linear Correlation Coefficient (CC), AUC-Judd and shuffled AUC. For SOD tasks, three standard metrics, PR-curve, F-measure, and MAE are used for evaluation, see [34].

We evaluated the fixed prior map generated by SOSA-Net, and compare it with 11 highly-recognized fixed models, including four classical models: ITTI [35], GBVS [36], AIM [37], BMS [38] and six deep-based learning models: eDN [39], SALICON [40], SU [41], Mr-CNN [42], Shallow-Net [43] and Deep-Net [44], and AS-Net [20]. Results are reported through PASCAL-S and MIT 1003 datasets. The results are shown in Table 2. SOSA-Net performs better than previous non-deep learning models and is competitive with current best performing deep learning competitors.

As shown in Table 2, SOSA-Net-14x14 performs better than previous non-in-depth learning models and is competitive compared with its currently best-performing in-depth learning competitors. Due to the relatively simple network structure and smaller output resolution, the proposed

TABLE 2. Quantitative comparison of different FP models on the PASCAL-S &MIT1003 dataset.

Methods	AUC-Judd		SIM		shuffled AUC		CC		NSS	
	PASCAL-S	MIT-1003	PASCAL-S	MIT-1003	PASCAL-S	MIT-1003	PASCAL-S	MIT-1003	PASCAL-S	MIT-1003
Mr-CNN	0.81	0.8	0.29	0.35	0.72	0.73	0.41	0.38	1.38	1.36
SALICON	0.87	0.85	0.43	0.42	0.69	0.74	0.61	0.53	1.81	1.86
Shallow-Net	--	--	--	--	0.71	0.68	--	--	1.91	1.6
Deep-Net	0.87	0.86	0.42	0.4	0.73	0.71	0.58	0.52	1.73	1.72
SU	--	--	--	--	0.73	0.73	--	--	2.21	2.08
eDN	--	0.85	--	0.3	0.65	0.66	--	0.41	1.41	1.29
BMS	--	0.79	--	0.33	0.68	0.69	--	0.36	1.32	1.25
AIM	0.77	0.79	0.28	0.27	0.66	0.68	0.34	0.26	1.01	0.82
GBVS	0.83	0.82	0.38	0.36	0.65	0.66	0.47	0.42	1.35	1.38
AS-Net	0.9	0.88	0.6	0.52	0.73	0.75	0.71	0.65	2.28	2.3
ITTI	0.84	0.77	0.41	0.32	0.64	0.66	0.42	0.33	1.33	1.11
SOSA-Net-14× 14	0.89	0.91	0.68	0.58	0.77	0.75	0.69	0.68	2.31	2.29
SOSA-Net-28 × 28	0.92	0.93	0.71	0.62	0.79	0.77	0.71	0.69	2.37	2.33

SOSA-Net has significant advantages. In addition, SOSA-Net-28x28 produces better results, which indicates that the proposed SOSA-Ne may obtain better FP results when more detailed spatial information is taken into consideration.

Here, we evaluate the performance of SOSA-Net on its main task: SOD. We quantitatively studied three widely used datasets and a self-constructed dataset, namely ECCSD [44], HKU-IS, PASCAL-S and MOD3. We compare SOSA-Net with alternatives based on in-depth learning: LEGS [45], MDF [26], DS [46], SU [41], DCL [47], ELD [48], RFCN [49], DHS [50], HEDS [51], NLDF [52], DLS [53], AMU [54], UCF [55], SRM [56], RC [57], MC [58], DSSOD [59], RSD [60], AS-Net. We also consider several classical non-in-depth learning models: HS [44], DRFI [61] and wCtr [62], DSR [63], and CHM [64]. Their results are provided by authors or by the running of their open source implementations through the original settings. In the following table 3, we report the maximum F measurements of Fβ and MAE scores. Overall, the proposed method achieves better performance on three data sets using all evaluation indicators. The qualitative results of the sample images from the above datasets show that the proposed SOSA-Net is very suitable for various complex scenarios.

The P-R curve is used to compare the proposed method in this article with the existing one. In Figure 5, we depict PR curves generated by our method and the most advanced previous methods on three popular datasets HKU-IS (blue dotted frame), KITTI (yellow dotted frame), PASCAL-S (red dotted frame) and self-made MOD3 data set (green dotted frame). Obviously, by synthesizing the results of multiple data sets, the algorithm proposed in this article can achieve the best results. We can also find that when the recall score is

TABLE 3. The F-measure and MAE scores of SOD on 5 different datasets.

Methods	PASCALS		HKU-IS		ECCSD		MSRA-B		MOD3	
	Fβ	MAE	Fβ	MAE	Fβ	MAE	Fβ	MAE	Fβ	MAE
DRFI	0.812	0.149	0.819	0.131	0.801	0.255	0.831	0.089	0.736	0.279
HS	0.636	0.259	0.711	0.213	0.729	0.223	0.821	0.163	0.723	0.284
wCtr	0.611	0.193	0.694	0.138	0.672	0.178	0.731	0.182	0.591	0.312
RFCN	0.862	0.127	0.879	0.081	0.816	0.235	0.942	0.088	0.824	0.229
RC	0.631	0.245	0.716	0.185	0.698	0.276	0.805	0.167	0.619	0.375
LEGS	0.749	0.155	0.812	0.101	0.831	0.119	0.847	0.125	0.718	0.326
CHM	0.611	0.275	0.696	0.205	0.678	0.289	0.792	0.179	0.696	0.383
DSR	0.627	0.255	0.703	0.235	0.692	0.279	0.811	0.168	0.618	0.379
SU	0.77	0.12	--	--	0.88	0.061	0.817	0.142	0.735	0.228
ELD	0.768	0.121	0.843	0.073	0.816	0.168	0.913	0.048	0.824	0.208
AMU	0.834	0.098	0.918	0.052	0.889	0.058	0.906	0.097	0.689	0.352
MDF	0.765	0.147	0.859	0.131	0.798	0.169	0.882	0.125	0.767	0.265
MC	0.723	0.149	0.792	0.103	0.732	0.185	0.875	0.064	0.698	0.325
AS-Net	0.857	0.072	0.92	0.035	0.928	0.043	0.932	0.049	0.825	0.098
DSSOD	0.83	0.08	0.913	0.039	0.896	0.045	0.934	0.055	0.833	0.196
RSD	0.848	0.083	0.896	0.033	0.931	0.047	0.918	0.043	0.846	0.081
SOSA-Net	0.916	0.078	0.935	0.028	0.935	0.039	0.949	0.042	0.851	0.076

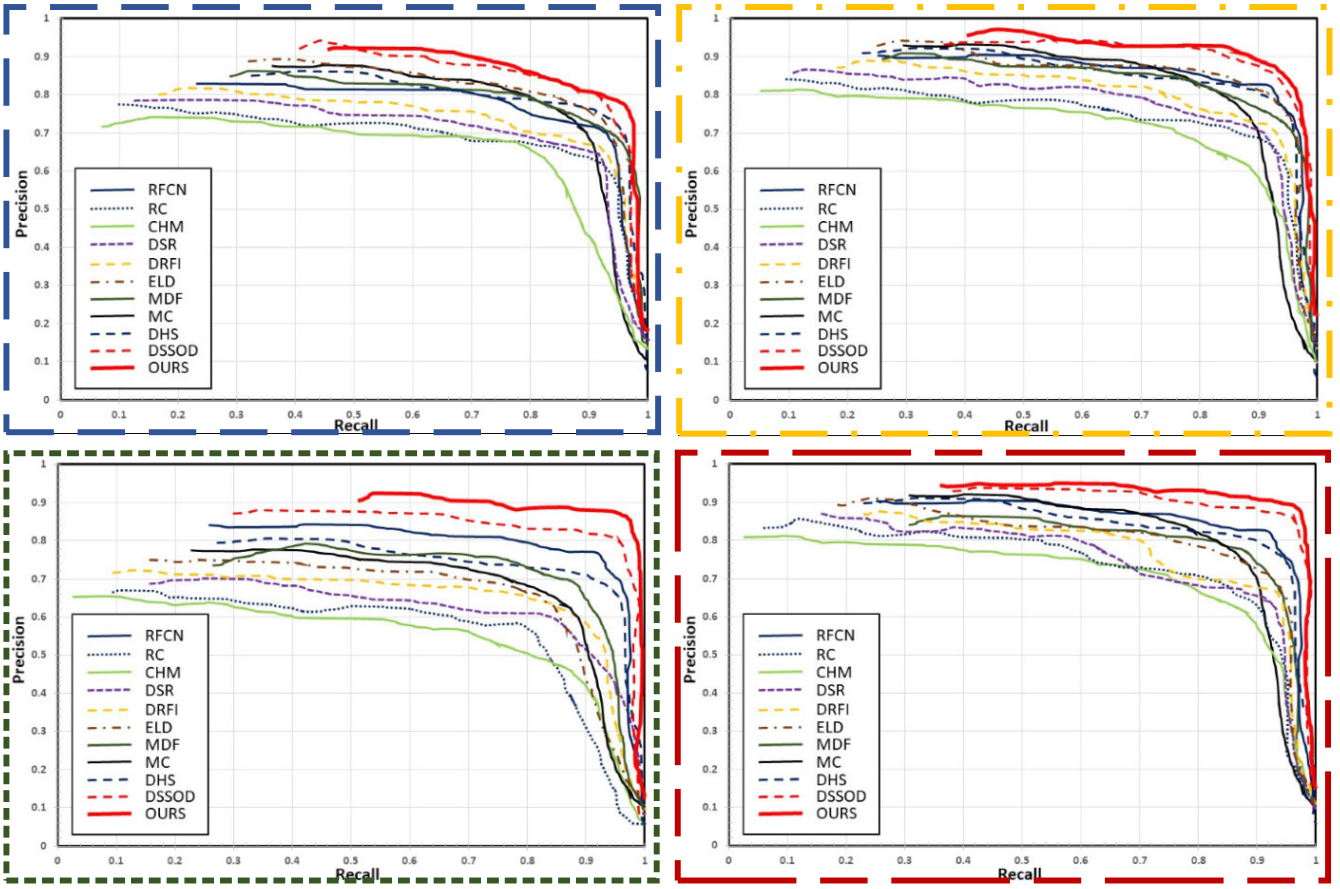


FIGURE 5. SOD results with P-R curve on three widely used benchmarks.

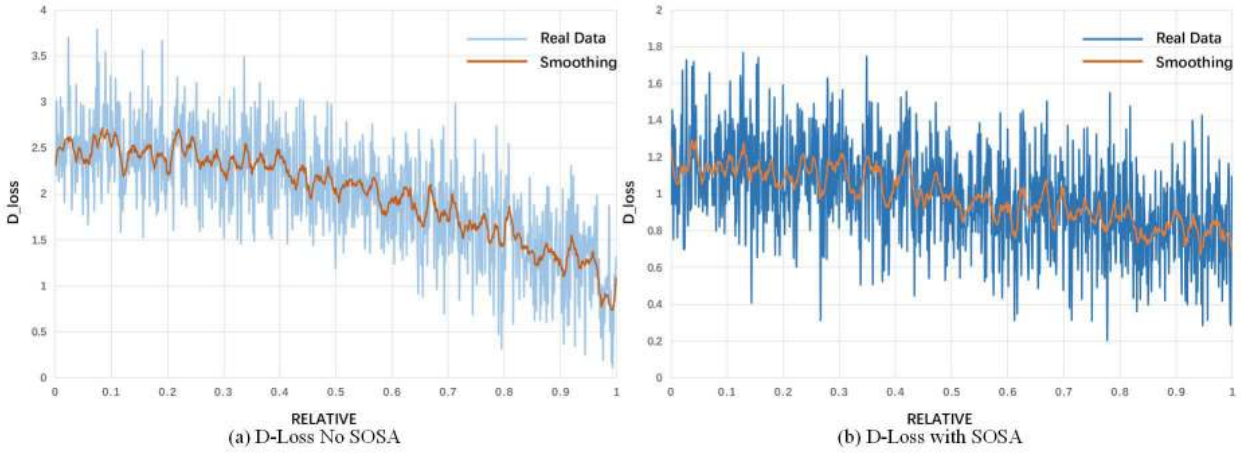


FIGURE 6. D-loss trend with and without SOSA module.

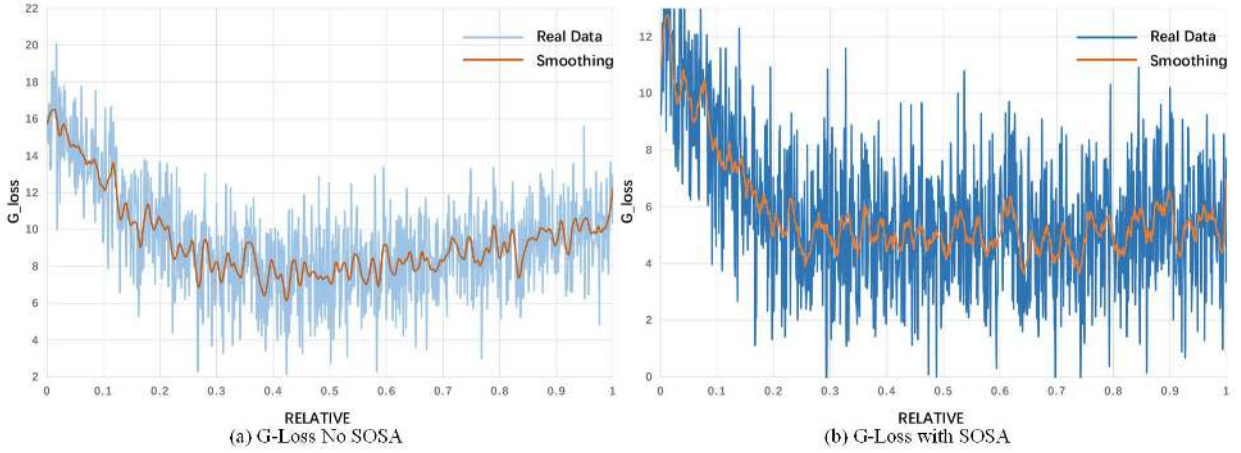


FIGURE 7. G-loss trend with and without SOSA module.

close to 1, the accuracy of our method is much higher, which reflects that our false alarm is much lower than that of other methods, indicating the effectiveness of our strategy, and the visual attention map thus obtained looks closer to the basic facts.

In order to fully verify the validity of the generated model, a 1:1 balanced update generation model and detection model were used. In the experiment, the loss function is optimized by the Adam algorithm, and the learning rate of the discriminator is set to 0.0003. The learning rate of the generator is set to 0.0001, the momentum is 0.5, and each batch is 128 samples, the model reached convergence after 10,000 iterations. Figure 6 and Figure 7 show the trends of discriminator loss function (D-loss) and generator loss function (G-loss) during the training process using the MOD1 data set combined with the SOSA module and the VSA-CGAN model without the SOSA module. The loss function is relatively smooth in the early stage of training, and the undulating fluctuation is more obvious in the later stage. However, in general, the discriminator loss function gradually decreases, and the generator loss function first drops rapidly and then rises slowly. It is verified that the model has a faster convergence speed and the overall quality of the model is higher.

Figure 8 shows the loss function of real data passing through the discriminator (D_loss_real) and the loss function of generated data passing through the discriminator (D_loss_fake). As the number of iterations increases, they all show a gradual downward trend.

C. COMPARISON WITH OTHER ADVANCED TECHNOLOGY

The Celeb A dataset is a large-scale facial feature dataset with 40 attribute tags per image (eg ''male'', ''eyes'', ''beard'', ''bangs'', etc.). Single Image Super Resolution Reconstruction (SISR) is a challenging task for computer vision and machine learning that attempts to reconstruct high resolution (HR) images from corresponding portions of low resolution (LR). On the CelebA face dataset, we have qualitative comparisons of the SISR task generation effects with several currently recognized SR advanced generation models: CA-GAN [11], CF-GAN [12], bicubic [65], pix2pix [66], SRGAN [67] and FCGAN [68].

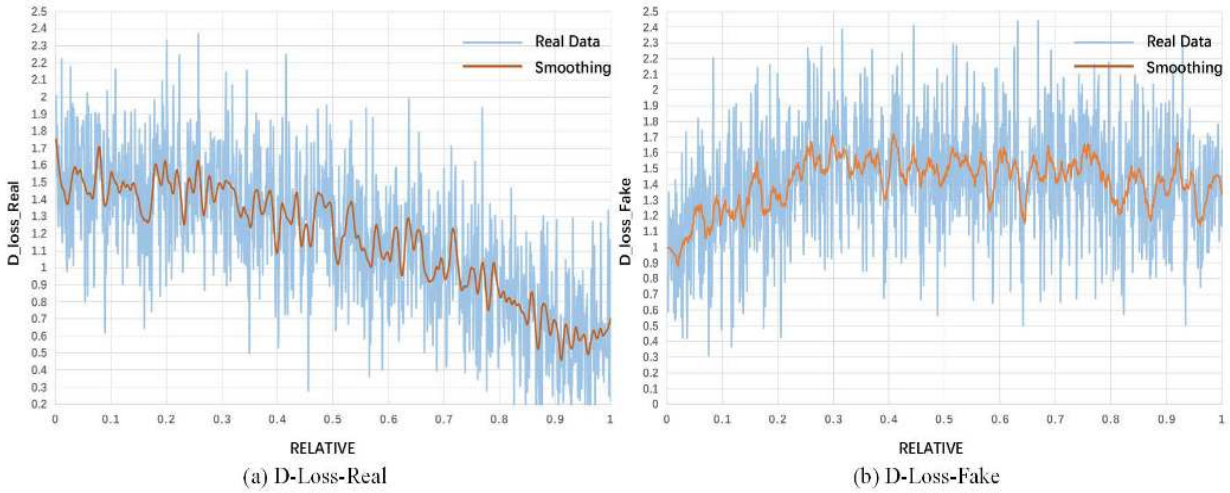


FIGURE 8. Real sample-the trend of loss function of generated sample in discriminator.

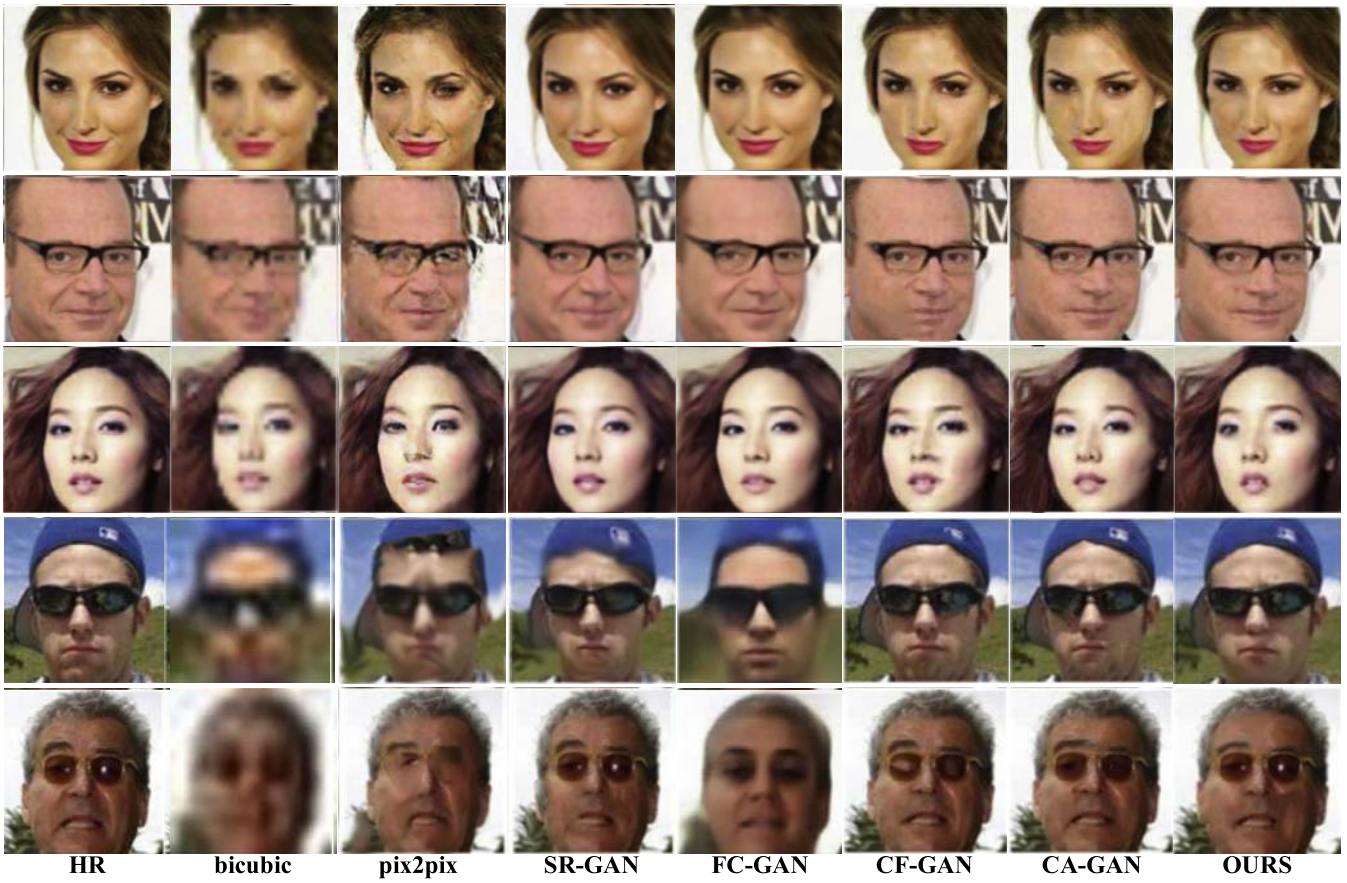


FIGURE 9. Different model generation results on the CelebA dataset.

The experiment iteratively trains 50 epochs on the Celeb A dataset, each of which has 10,000 iterations. The number of images for each batch of training is 64. The optimizer used is Adam, where the parameter $\beta_1 = 0.5$, $\beta_2 = 0.9$, and the initial learning rates of the generated network and the discriminant network are 0.0001 and 0.0004, respectively. The learning rate attenuation factor is 0.95.

First, we downsample the HR image (128×128) to a resolution of 32×32. Then, we use the bicubic interpolation algorithm to generate the interpolated image (I^B, size 128×128), and finally construct the I^B and I^{HR} images into the input

FIGURE 10. Different model generation results on the CelebA dataset.

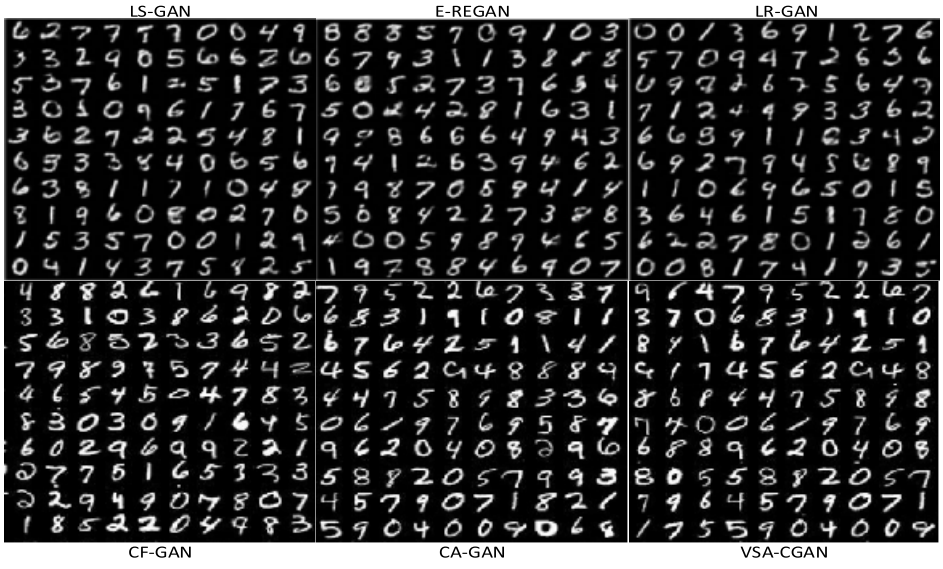


FIGURE 11. MNIST handwritten data set training results.

and output pairs $\left(I^{B}, I^{HR}\right)$. Therefore, the model's input and output images are the same size, 128 × 128, with three color channels. We report the qualitative results in the Figure 9. From the results, we can see that our method is obviously superior to other methods in performance, no matter facial expression, posture, lighting, occlusion (wearing glasses or hats), using this model can generate high-quality facial images more accurately, and reconstruct small local features.

We provide quantitative analysis of PSNR and SSIM evaluation indicators [68] in Table 4 and Table 5, and provide quantitative analysis of training results for different magnification ratios (2×, 3×, and 4×) in the Table 5. As can be seen from

FIGURE 12. Training results of the MOD3 data set.

TABLE 4. The F-measure and MAE scores of SOD on 5 different datasets.

Methods	Bicubic	Pix2Pix	SR-GAN	FC-GAN
PSNR	31.23	30.25	32.53	32.64
SSIM	0.775	0.738	0.859	0.863
Methods	CF-GAN	CA-GAN	OURS	
PSNR	34.11	33.07	33.15	
SSIM	0.862	0.883	0.872	

TABLE 5. The quantitative comparison with different upscaling factors.

Ratio	G-Loss	D-Loss	PSNR	SSIM
2 ×	0.029	0.305	35.12	0.96
3 ×	0.038	0.306	34.01	0.89
4×	0.049	0.309	32.71	0.88

TABLE 6. Comparisons of detection effect of multiple model objects.

Method	Dataset	mAP(%)	P (%)
Faster R-CNN [74]	KITTI	72.61	65.22
	MOD2	53.82	38.63
DSOD300 [75]	KITTI	78.69	71.68
	MOD2	56.05	42.32
DSSD513 [76]	KITTI	80.11	79.42
	MOD2	59.16	49.79
YOLOv3 [73]	KITTI	82.86	83.62
	MOD2	63.59	56.86
VSA-CGAN+L2 SVM	KITTI	82.83	83.77
	MOD2	73.36	66.05

the results, our method has significant advantages compared to other methods.

Figure 10 is the result of the generation of different conditional features on the model VSA-CGAN. (a) is the generated image of the conditional feature of the hat, (b) is the generated image of the male and black hair conditional features, (c) is the generated image of the female and blond conditional features, and (d) is the conditional feature of the added glasses. Generated image. Observing the results, we can find that the sample details generated by VSA-CGAN are clear, and the contour features of the facial features are obviously different at the same time.

MNIST data set is a grayscale picture with 70000 handwritten numbers, including 60000 training samples and 10000 test samples. Each picture is 28×28 pixels in size, with a total of 10 categories of 0-9. Figure 11 shows the generation effect diagram of VSA-CGAN model and several mainstream generation models LS-GAN [69], E-REGAN [70], LR-GAN [71] by adding 0-9 different conditional features to the MNIST data set, the resolution of the generated image is 128×128. The samples generated by conditional features

TABLE 7. Comparison of data augmentation effects of various models.

Augmentation Method	mAP(%)		P (%)		IS		FID	
	KITTI	MOD2	KITTI	MOD2	KITTI	MOD2	KITTI	MOD2
YOLOv3	60.16	49.28	50.35	40.96	/	/	/	/
Copy	60.23	49.37	50.39	41.03	/	/	/	/
Affine Transformation	60.75	49.78	51.06	41.12	/	/	/	/
SA-GAN	65.37	55.18	55.68	45.36	28.28	19.27	39.78	57.25
SN-GAN	64.27	54.43	54.77	44.13	23.17	14.96	43.53	60.16
LS-GAN	62.47	51.98	52.39	41.96	19.28	12.52	45.73	62.35
E-REGAN	64.58	54.79	54.96	44.75	26.35	17.43	40.91	59.43
CF-GAN	66.87	58.87	57.02	46.98	28.76	22.54	35.23	37.11
CA-GAN	68.23	57.08	56.38	47.81	28.53	25.76	39.11	39.87
VSA-CGAN	67.32	60.12	56.12	49.28	27.98	28.96	38.27	39.13

are iterated for 10 times in total. After only one iteration, the model learns very accurate features. When it reaches 10 iterations, the model generates sample wheel in this article. It has the clearest outline and reasonable structure, which is almost identical with the real handwritten digital image.

Figure 12 shows part of the simulation results of VSA-CGAN on MOD3 data set, and the generated sample resolution is 256 × 256. From the experimental results, the shape and structure distribution characteristics of various military objects are similar to the real samples, such as: tank turret, track, camouflage, helicopter wing and cabin, attached weapons, helicopter propeller radiation, etc., which are fully learned. The resampling of the data distribution in the generation process has obtained samples with rich texture and diverse posture, such as different shapes of tank turret and body, different posture of fuselage during helicopter flight, propeller of helicopter rotation, different background environment, etc., resulting in new sampling that the training samples do not have.

In order to evaluate the performance of the simulation method represented by VSA-CGAN model, the network structure is first used to extract the characteristics of test data, and then the feature vectors after flattening the connection are applied to the linear model. Finally, the performance of the linear model is evaluated. In this experiment, the convolution of each layer in the identification network D will undergo the following procedure: after the training is completed, their label part of the connection will be removed, their features are acquired by max pooling, and the features of each layer are flattened and connected to output multidimensional feature vectors, which are further applied to Linear Support Vector Machine (LSVM) to complete the classification. In addition, on the same test data set, several current mainstream object detection models based on in-depth learning are used to conduct comparative experiments. We choose the first picture set of KITTI dataset [72].

Average precision (AP) is the most intuitive criterion for evaluating the accuracy of in-depth learning detection model. AP measures the accuracy of detection algorithm from two angles of recall and accuracy, and can be used to analyze the detection effect of a single category. Mean AP (mAP) is the average of each category of AP. The higher the mAP, the higher the comprehensive performance of the model in all categories. At the same time, in order to verify the validity of the model for multi-object detection, we add the evaluation index detection rate [72].

The performance comparison of each classification method is shown in Table 6. In KITTI traffic scene data set, the detection rate of the five methods is more than 65%. Among them, YOLOv3 [73] has the best detection accuracy, but the difference between this model and the other three models is small, and it is obviously better than the other three models. In MOD2 self-sufficient data set, there are only a small number of samples, and the object and background are complex, so the detection accuracy of the other four kinds of models decreases obviously. On the contrary, due to the reasonable strategy setting, the model proposed in this article can achieve detection accuracy, thus making it better than the other four types of models. The experimental results validate that the model can learn the essential features of the image and has a great ability to imitate multi-distributed data. When applied to the simulation of specific object images, it can generate specific types of samples efficiently and randomly.

YOLOv3 model is outstanding in image classification experiments, but its parameters are numerous, so when the number of training samples is small, it is easy to produce overfitting. Data Augmentation can increase the number of samples and improve network performance. We use ESA-CGAN synthetic image as a means of data enhancement and compare it with affine transformation so as to verify the validity of synthetic samples. We construct a three-layer conv + pooling structure, followed by two hidden FC layers, and finally direct

the output to soft-max to get the prediction probability of each category. We then select 300 original images of each category from two datasets, synthesize 150 images by adding data enhancement method, and form the training set of classifier network. In addition, 200 original samples are selected as test sets for comparative experiments.

Data enhancement is data enlargement, and its aims are to increase the size of data sets and prevent model overfitting in machine learning. We also conducted comparative experiments to understand what is the gain of data expansion tasks achieved by the combination of SOSA-Net and the cooperative strategy with migration learning as proposed in this article. As the experimental results show in the Table 7, affine transformation mainly includes three modes: flip, translation and zoom. On KITTI dataset, because the CNN classifier has achieved high accuracy, each of the data enhancement method can hardly achieve conspicuous improvement on the performance; on MOD2 dataset, the classification performance of ESA-CGAN enhanced dataset is better than that of duplicate samples and CNN classifier that has been enhanced by affine transformation, which indicates that it has avoided overfitting and that the synthesized images generated by the method proposed in this article prove to be valid.

V. CONCLUSION

Based on the latest GAN research progress, this article proposes a new generated confrontation network model for image data simulation application problem–the conditional generation confrontation network model (VSA-CGAN) of fusion visual perception and self-attention mechanism, the number and label of training samples under the conditional constraints, samples of the expected category can be generated, and the designed enhancement object self-focus fusion mechanism SOSA-Net can efficiently extract object features which ensures high quality sample output. Meanwhile, the sample generation capability of the model is verified on the standard datasets MNSIT and Celeb A and the self-constructed dataset MOD. From the visual point of view, the generated sample results have a similar structure and color distribution as the real image, the texture information is random, and the background noise is better suppressed at the same time. In addition, the features extracted by the model are used in the image object detection experiment, which has obvious advantages compared with the performance of other detection algorithms, and effectively validates the model's ability to imitate the essential features of the image object. This article does not simply sample or transform an existing image to achieve the purpose of data augmentation. Instead, it can fully understand the characteristics of a specific object image through focused learning, and then generate a new object image, compared to the traditional object. The simulation augmentation method not only improves the efficiency of generating multi-distributed data, but also effectively improves the quality of generation. The next step will be to conduct in-depth research on how to further improve the quality of the details of the generated object.

REFERENCES

[1] L. J. Goodfellow, "Generative adversarial nets," in *Proc. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

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

[3] Z. Li, W. Zhang, L. Wang, A. Cai, N. Liang, B. Yan, and L. Li, "A sinogram inpainting method based on generative adversarial network for limited-angle computed tomography," 2019, *arXiv:1903.03984*. [Online]. Available: http://arxiv.org/abs/1903.03984

[4] T. M. Quan, T. Nguyen-Duc, and W.-K. Jeong, "Compressed sensing MRI reconstruction using a generative adversarial network with a cyclic loss," *IEEE Trans. Med. Imag.*, vol. 37, no. 6, pp. 1488–1497, Jun. 2018.

[5] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *Comput. Sci.*, Jan. 2015.

[6] K. M. He, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Oct. 2014.

[7] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," 2018, *arXiv:1805.08318*. [Online]. Available: http://arxiv.org/abs/1805.08318

[8] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," 2018, *arXiv:1809.11096*. [Online]. Available: http://arxiv.org/abs/1809.11096

[9] Y. Feng, C. Zhang, B. Qiang, Y. Zhang, and J. Shang, "GP-WIRGAN: Wasserstein image cycle generation confrontation network model optimized by gradient penalty," *J. Comput. Sci.*, May 2019.

[10] Y. Ji and L. Ma, "Self attention generative adversarial network under condition constraints," Xi'an Univ., Xi'an, China, Tech. Rep., 2019.

[11] J. Feng, X. Feng, J. Chen, X. Cao, X. Zhang, L. Jiao, and T. Yu, "Generative adversarial networks based on collaborative learning and attention mechanism for hyperspectral image classification," *Remote Sens.*, vol. 12, no. 7, p. 1149, Apr. 2020.

[12] M. Chen, Z. Liu, L. Ye, and Y. Wang, "Attentional coarse-and-fine generative adversarial networks for image inpainting," *Neurocomputing*, vol. 405, pp. 259–269, Sep. 2020, doi: 10.1016/j.neucom.2020.03.090.

[13] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using CNN with attention mechanism," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2439–2450, May 2019.

[14] Y. Yan, J. Ren, G. Sun, H. Zhao, J. Han, X. Li, S. Marshall, and J. Zhan, "Unsupervised image saliency detection with gestalt-laws guided optimization and visual attention based refinement," *Pattern Recognit.*, vol. 79, pp. 65–78, Jul. 2018.

[15] X. Hua, X. Wang, T. Rui, D. Wang, and F. Shao, "Real-time object detection in remote sensing images based on visual perception and memory reasoning," *Electronics*, vol. 8, no. 10, p. 1151, Oct. 2019.

[16] C. Luo, L. Jin, and Z. Sun, "A multi-object rectified attention network for scene text recognition," *Pattern Recognit.*, vol. 90, pp. 109–118, Jun. 2019.

[17] K. J. Mintz, Y. Zhou, and R. M. Leblanc, "Recent development of carbon quantum dots regarding their optical properties, photoluminescence mechanism, and core structure," *Nanoscale*, vol. 11, no. 11, pp. 4634–4652, 2019.

[18] H. Li, C. Li, and J. Ren, "Attention mechanism improved convolution neural network for remote sensing image object detection," *J. Image Graph.*, pp. 1400–1408, 2019.

[19] N. Kodali, J. Abernethy, J. Hays, and Z. Kira, "On convergence and stability of GANs," 2017, *arXiv:1705.07215*. [Online]. Available: http://arxiv.org/abs/1705.07215

[20] W. Wang, "Salient Object Detection Driven by Fixation Prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.(CVPR)*, vol. 2018, pp. 1711–1720.

[21] Y. Qin, N. Mitra, and P. Wonka, "How does lipschitz regularization influence GAN training?" 2018, *arXiv:1811.09567*. [Online]. Available: http://arxiv.org/abs/1811.09567

[22] O. Ronneberger, P. Fischer, and B. Thomas, "U-Net: Convolutional Networks for Biomedical Image Segmentation," 2015, arXiv:1505.0 4597. [Online]. Available: https://arxiv.org/abs/1505.04597

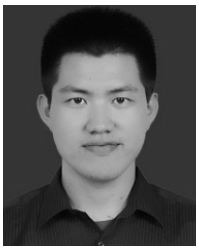
[23] C. Li and W. Michael, "Precomputed real-time texture synthesis with Markovian generative adversarial networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 702–716.

[24] L. Deng, "The MNIST database of handwritten digit images for machine learning research best of the Web," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 141–142, Nov. 2012.

[25] Z. Liu, P. Luo, X. Wang, and X. Tang. *Large-Scale Celebfaces Attribu Tes (CelebA) Dataset*. Accessed: Jul. 20, 2017. [Online]. Available: http://mmlab.ie.cuhk.ed u.hk/projects/CelebA.html
[26] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5455–5463.
[27] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, Sep. 2013.
[28] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 280–287.
[29] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 2016–2113.
[30] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," 2018, *arXiv:1802.05957*. [Online]. Available: http://arxiv.org/abs/1802.05957
[31] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," 2016, *arXiv:1606.03498*. [Online]. Available: http://arxiv.org/abs/1606.03498
[32] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," 2017, *arXiv:1706.08500*. [Online]. Available: http://arxiv.org/abs/1706.08500
[33] J. Xie, "Infrared target simulation method based on the generation of antagonism neural network," *Acta. Optica. Sinica.*, vol. 39, no. 3, p. 1, Oct. 2019.
[34] W. Wang, J. Shen, L. Shao, and F. Porikli, "Correspondence driven saliency transfer," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5025–5034, Nov. 2016.
[35] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, 1998.
[36] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 19, 2007, pp. 545–552.
[37] N. D. B. Bruce and J. K. Tsotsos, "Saliency, attention, and visual search: An information theoretic approach," *J. Vis.*, vol. 9, no. 3, p. 5, Mar. 2009.
[38] J. Zhang and S. Sclaroff, "Saliency detection: A Boolean map approach," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 153–160.
[39] E. Vig, M. Dorr, and D. Cox, "Large-scale optimization of hierarchical features for saliency prediction in natural images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2798–2805.
[40] X. Huang, C. Shen, X. Boix, and Q. Zhao, "SALICON: Reducing the semantic gap in saliency prediction by adapting deep neural networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 262–270.
[41] S. S. S. Kruthiventi, "Saliency unified: A deep architecture for simultaneous eye fixation prediction and salient object segmen tation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Dec. 2016, pp. 5781–5790.
[42] N. Liu, J. Han, and X. Li, "Learning to predict eye fixations via multiresolution convolutional neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 2, pp. 392–404, Feb. 2018.
[43] J. Pan, E. Sayrol, X. Giro-I-Nieto, K. McGuinness, and N. E. OConnor, "Shallow and deep convolutional networks for saliency prediction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 598–606.
[44] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1155–1162.
[45] L. Wang, H. Lu, X. Ruan, and M.-H. Yang, "Deep networks for saliency detection via local estimation and global search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3183–3192.
[46] X. Li, L. Zhao, L. Wei, M.-H. Yang, F. Wu, Y. Zhuang, H. Ling, and J. Wang, "DeepSaliency: Multi-task deep neural network model for salient object detection," *IEEE Trans. Image Process.*, vol. 25, no. 8, pp. 3919–3930, Aug. 2016.
[47] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 478–487.
[48] G. Lee, Y.-W. Tai, and J. Kim, "Deep saliency with encoded low level distance map and high level features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 660–668.
[49] L. Wang, "Saliency detection with recurrent fully convolutional networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 2016, pp. 825–841.
[50] N. Liu and J. Han, "DHSNet: Deep hierarchical saliency network for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 678–686.
[51] J. Zhang, Y. Dai, F. Porikli, and M. He, "Multi-scale salient object detection with pyramid spatial pooling," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Dec. 2017, pp. 1286–1291.
[52] Z. Luo, A. Mishra, A. Achkar, J. Eichel, S. Li, and P.-M. Jodoin, "Non-local deep features for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6609–6617.
[53] P. Hu, B. Shuai, J. Liu, and G. Wang, "Deep level sets for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2300–2309.
[54] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating multi-level convolutional features for salient object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 202–211.
[55] P. Zhang, D. Wang, H. Lu, H. Wang, and B. Yin, "Learning uncertain convolutional features for accurate saliency detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 212–221.
[56] T. Wang, A. Borji, L. Zhang, P. Zhang, and H. Lu, "A stagewise refinement model for detecting salient objects in images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4019–4028.
[57] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, Mar. 2015.
[58] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1265–1274.
[59] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. S. Torr, "Deeply supervised salient object detection with short connections," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 815–828, Apr. 2019.
[60] Y. Wang, X. Wei, L. Ding, X. Tang, and H. Zhang, "A robust visual tracking method via local feature extraction and saliency detection," *Vis. Comput.*, vol. 36, no. 4, pp. 683–700, Apr. 2019.
[61] J. Wang, H. Jiang, Z. Yuan, M.-M. Cheng, X. Hu, and N. Zheng, "Salient object detection: A discriminative regional feature integration approach," *Int. J. Comput. Vis.*, vol. 123, no. 2, pp. 251–268, Jun. 2017.
[62] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2814–2821.
[63] X. Li, H. Lu, L. Zhang, X. Ruan, and M.-H. Yang, "Saliency detection via dense and sparse reconstruction," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 2976–2983.
[64] X. Li, Y. Li, C. Shen, A. Dick, and A. V. D. Hengel, "Contextual Hypergraph Modeling for Salient Object Detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2013, pp. 3328–3335.
[65] Z. Roman, E. Michael, and P. Matan, "On single image scale-up using sparse-representations," in *Proc. Int. Conf. Cur. Surf.*, 2010, pp. 711–730.
[66] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-Image translation with conditional adversarial networks," 2016, *arXiv:1611.07004*. [Online]. Available: http://arxiv.org/abs/1611.07004
[67] C. Ledig, L. Theis, F. Huszar, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 105–114.
[68] X. Tang, Z. Wang, W. Luo, and S. Gao, "Face aging with identity-preserved conditional generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7939–7947.
[69] G. Qi, "Loss-sensitive generative adversarial networks on Lipschitz densities," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Oct. 2017, pp. 1118–1140.
[70] G. Wang, J. Qiao, and L. Wang, "A generative adversary network in the sense of energy function," *J. Aut. Chemis.*, vol. 44, no. 5, 2018, pp. 28–38.
[71] J. Yang, "LR-GAN: Layered recursive generative adversarial networks for image generation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 1–7.
[72] X. Wang, X. Hua, F. Xiao, Y. Li, X. Hu, and P. Sun, "Multi-object detection in traffic scenes based on improved SSD," *Electronics*, vol. 7, no. 11, p. 302, Nov. 2018.
[73] R. Joseph and A. Farhadi, "YOLOv3: An incremental improvemen," Tech. Rep., 2018.

[74] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," 2015, *arXiv:1506.01497*. [Online]. Available: http://arxiv.org/abs/1506.01497
[75] Z. Shen, Z. Liu, J. Li, Y.-G. Jiang, Y. Chen, and X. Xue, "DSOD: Learning deeply supervised object detectors from scratch," 2017, *arXiv:1708.01241*. [Online]. Available: http://arxiv.org/abs/1708.01241
[76] C. Fu, "DSSD : Deconvolutional single shot detector," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 1–11.

PENG ZHANG received the bachelor's degree in engineering from the East China University of Science and Technology, Shanghai, China, in 2018. He is currently pursuing the master's degree with the Army Engineering University of PLA. His main research interests include computer vision, intelligence, and 3-D point cloud processing.



XIA HUA received the M.S. degree from the Army Engineering University of PLA, where he is currently pursuing the Ph.D. degree. His main research interests include computer graphics, machine vision, digital image processing, and artificial intelligence.

XINQING WANG received the Ph.D. degree (Hons.). He is currently a Professor and a Doctoral Tutor with the Army Engineering University of PLA. His main research interests include electromechanical control, intelligent signal processing, and machine vision.

TING RUI (Senior Member, IEEE) received the M.S. and Ph.D. degrees from the PLA University of Science and Technology, Nanjing, China, in 1998 and 2001, respectively. He is currently a Professor with the Army Engineering University of PLA. He has authored or coauthored more than 80 scientific articles. His research interests include computer vision, machine learning, multimedia, and video surveillance.

HAITAO ZHANG received the Ph.D. degree (Hons.). He is currently a Professor and a Doctoral Tutor with the Army Engineering University of PLA. His main research interests include electromechanical control and intelligent signal processing.

FAMING SHAO was born in 1978. He received the Ph.D. degree from the Army Engineering University of PLA, China, in 2020. He is currently an Associate Professor with the Army Engineering University of PLA. His research interests include signal processing, deep learning, and software engineering.

DONG WANG received the Ph.D. degree. He is currently a Lecturer with the Army Engineering University of PLA. His main research interests include computer power control, intelligent signal processing, and artificial intelligence.