Generative deep deconvolutional neural network for increasing and diversifying training data

Runnan Qin, Rui Wang
Key Laboratory of Precision Opto-mechatronics Technology, Ministry of Education, School of Instrumentation Science and Opto-electronics Engineering, Beihang University
Beijing 100191, China
E-mail: QinRunnan@buaa.edu.cn, wangr@buaa.edu.cn

***Abstract*—Large amount of annotated images with rich variations are needed to train a deep network for detecting instance object in unstructured environment. Addressing the problem that the artificial acquisition and manual annotation is time-consuming, the generative deep deconvolutional neural network (GDDNE) to increase and diversify training data through the supervised learning strategy is created in this paper. Specifically, our network can not only generate with different styles such as shift, zoom, brightness and other superimposed transformations, but also interpolate generate the new ones between given viewpoints images in training samples. With 180 viewpoints realistic images in training samples: 30 rotation angles in plane and 6 angles of depression, our network can finally generated 1000 diversified viewpoint images and 21 kinds of data transformations for each instance object. Abundant experiments demonstrate that the remarkable performance of our generative network used in the generation task of large magnitude.**

Keywords—image generation, deconvolutional neural network, training data.

I. INTRODUCTION

The diversified training data – large amount of annotated images with rich variations are needed to train a deep network for detecting in unstructured environment. Traditionally, such a massive work must be taken via realistic shooting images for instance objects in random background [1]. Nevertheless artificial acquisition and manual annotation in such way is expensive and time-consuming. Besides, many expansion works of training data such as data augmentation can only transform the illumination or saturation of images, but cannot change the posture information of instance objects in unstructured environment. Therefore, the generative model, which has the ability to generate new data samples by learning the joint probability distribution of the data samples and the labels, can complete the expansion work of different viewpoint images. In this paper, we use the generative model with end-to-end training strategy to learn a mapping relationship between the real images and the low dimension description such as viewpoints, then generate new images with designated description directly without complicated induction process.

According to different training strategies, generative models can be divided into three categories: supervised generation models, unsupervised generative models and semi-supervised generative models. In recent years, supervised generation models revalued again because of the acquisition of big data, the development of computer hardware, and the appearance of Convolutional Neural Network (CNN). Since the Deconvolutional Neural Network (DNN) [2] has been proposed for reconstruction of image features, an increasing amount of papers using such similar architecture to generate images. For examples, Tejas D. Kulkarni et al. [3] have designed the Deep Convolutional Inverse Graphics Network (DCIGN) to generate three-dimensional image of face; Dosovitskiy et al. [4][5] use a deconvolutional neural network to generate multi class chair images; The Deep Convolutional Generative Adversarial Networks (DCGAN) to produce high fidelity indoor images have designed by Alec Radford et al. [6] In addition, semi-supervised generative models and unsupervised generative models have also been widely developed. Usually, Semi-supervised generative models only need partially label samples, and the maximum expected algorithm is used to estimate parameters, which means that Gaussian Mixed Model (GMM) [7] and Hidden Markov Model (HMM) [8] or other maximum expected models can be used as a base classifier. As for unsupervised generative models, the prominent examples are Restricted Boltzmann Machine (RBM) [9] and Deep Boltzmann Machines (DBM) [10], which can achieve statistical modeling between its complex architecture and massive unlabeled dataset. However, semi-supervised generative models are usually applied to some small models, because they usually require a large amount of calculations during training process. And as compared with the supervised generative models, unsupervised generative models usually cannot accurately control the features of the generated images.

In what follows, we propose to use a kind of supervised generation model termed Generative Deep Deconvolutional Neural Network (GDDNE) to address the problem of increasing and diversifying training data, i.e. generating different viewpoints images and versatile data augmentation changes with limited key realistic samples in the training process. Difference from the generative works mentioned above, we not only just generate different images with the network, but also concentrate on how to guarantee the magnitude and the accuracy of the generated images to satisfy our specific application task for building a large-scale training

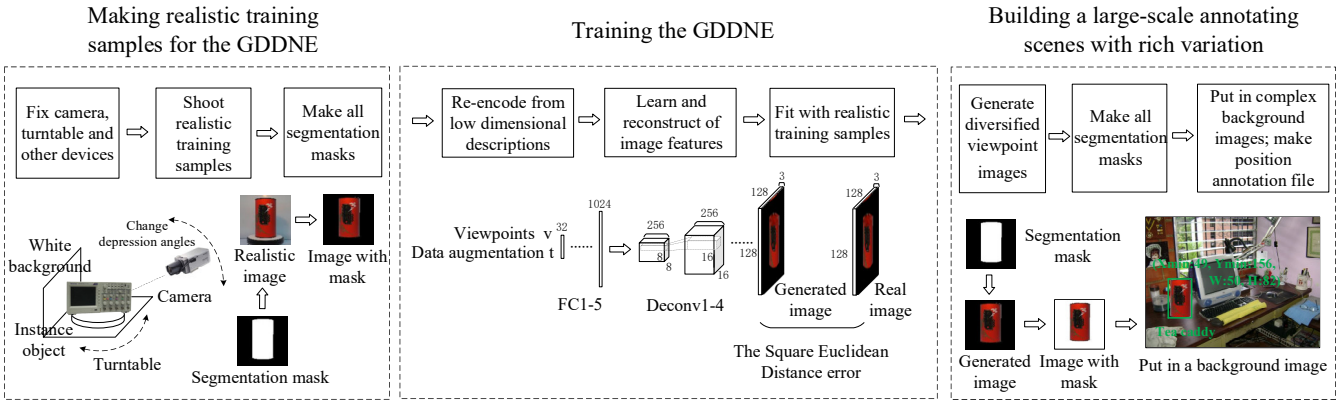

Fig. 1. The Overall architecture of the GDDNE system.

data with labels used in a deep network. Moreover, the images generated with supervised generation models [3]-[6] are one class of object such as human face, which may cause interference or category change phenomenon in the process of network training due to the different between-class features. Conversely, our GDDNE system ensure the specific instance object images generated, which effectively reducing the complexity of training process and improving the accuracy of generated images. The following will elaborate our GDDNE system including the realistic samples information, network architectures, network training and performance, as well as experimental results and analysis.

II. PROPOSED METHOD

To build annotated scenes with rich variation through the limited realistic labeled images, our GDDNE system is a three stage pipeline shown in Fig.1: 1) make the realistic samples information of specified instance objects (the labels and the key realistic images for training the GDDNE, the ground truth images for experimental test); 2) train the GDDNE to automatically output the qualified viewpoint images with data augmentation; 3) make segmentation masks of the images generated to extract the only part of instance objects; pick them into complex background images and automatically record the annotation files.

A. Provide realistic samples information

We need to provide the key realistic training samples that contain diversified viewpoint images of specified instance objects with labels to train the GDDNE, as well as supply the ground truth realistic samples to test in experiment. Thus, in order to complete the collection of diversified viewpoint images, firstly we put a turntable at constant speed in white background, then place instance objects on the turntable that rotate around its longitudinal axis. At the same time, a camera is installed at a fixed position, then take videos about instance objects that rotate 360 degrees with different depression angles of the camera. Each video sequence contains many realistic images, which can used as the ground truth realistic images.

Besides, we only select limited ground truth images as the key realistic training images for training the GDDNE, making the GDDNE can generate new viewpoint images between the limited given ones. In order to reduce the complexity of training process, we adjust the size of images to 128×128 pixel. What's more, we also make all segmentation masks (s) to extract the only part of images containing instance objects before input the key realistic training samples to the GDDNE for training, so as to avoid the interference of background and further optimize the training process of network.

Moreover, the labels in the key realistic training samples are the low dimensional descriptions of images, which consist of two vectors: v – the rotation angle and depression angle corresponding to the camera position (represented by their sine and cosine values); t – the parameters of data augmentation transformations. The randomly parameter vector t can gain six kinds of data augmentation changes: shift vertically or horizontally (0 to one-tenth of image size), zoom (100% to 135%), stretching vertically or horizontally (0 to one-tenth of image size), in-plane rotation (0 to 12 degrees), brightness (35% to 300%), saturation (35% to 300%). In training process, randomly parameter t are added to increase the variation of training samples and reduce overfitting.

B. The GDDNE architecture

The network structure of the GDDNE is shown in Fig.2. We give the key realistic training samples with the labels $D=\{(v^1, t^1),...,(v^N, t^N)\}$ and the key realistic training images $G=\{x^1_{RGB},...,x^N_{RGB}\}$ as the input of the GDDNE.

Then, three parts of the GDDNE structure are described in detail: firstly, a shared, re-encoding network to obtain the high dimensional hidden representation $h(v,t)$ from the input parameters D is built by fully connected layers FC1 to FC5. The two input vectors v and t are independently go through two layers of FC1 and FC2 with 512 neurons, then the outputs are concatenated with 1024 neurons. FC3 and FC4 both with 1024 neurons follow this independent processing, resulting in the output response of the layer FC5. Finally, FC5 outputs a 16384-dimensional vector and reshape the vector into 8x8 multichannel feature map.

Secondly, the reconstruction of image features is realized by the structure of three deconvolutional layers with 4x4 filters. And a convolutional layer with 3x3 filters follows each of them.

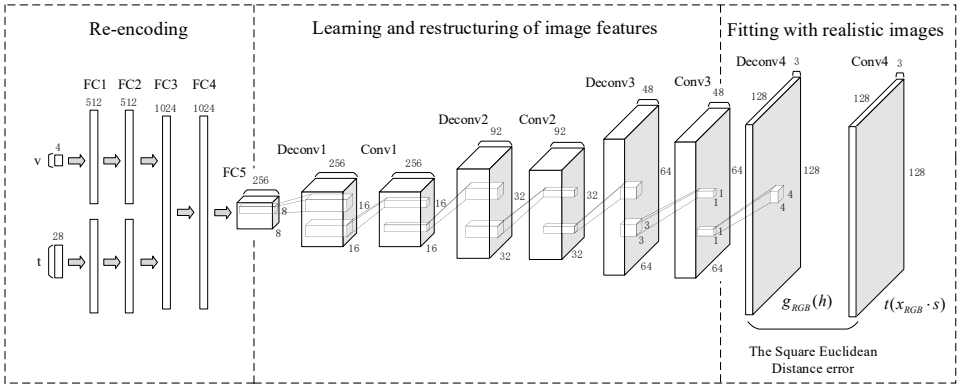

Fig. 2. The network structure of the GDDNE.

In order to reconstruct high dimensional images from 8x8 feature representation, we need to increase the space span of feature size, which is opposite to the pooling process in usual CNN. As illustrated in Fig.3, set up the parameters of each deconvolutional layer: kernel size=4, stride=2, pad=1, which can increases 2 times of the width and the height of the feature maps. Finally, the output of last deconv4 layer is 128x128 feature representation, which means the RGB image $i_{RGB}(h)$ has successfully predicted according to the high dimensional hidden representation $h(v,t)$.

In addition, a Rectified Linear Unit (ReLU) to accelerate the convergence rate of the network follows each layer in GDDNE, except the output layer.

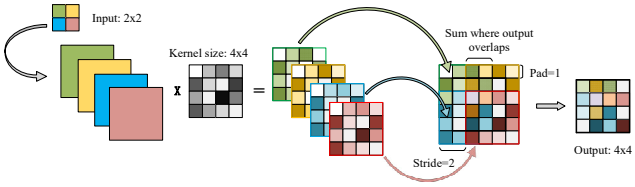

Fig. 3. Illustration of the calculation process of deconvolution.

Finally, to fit the generated images and the realistic images in training samples, minimizing the Squared Euclidean Distance (SED) between them that defined by formula (1).

$$SED = \sum_{i=1}^{N} \left\| g_{RGB}(h(v^i, t^i)) - t^i\left(x_{RGB}^i \cdot s^i\right) \right\|_2^2 \qquad (1)$$

C. *Network training details*

In order to training the GDDNE, we use the Caffe framework [11] of CNN to run on.

The network parameters W are consist of all layer weights and biases. It trained by minimizing the loss function, as shown in formula (2). In our experiments, the loss function is chosen as the SED error.

$$\min_W \sum_{i=1}^{N} L_{RGB}\left(g_{RGB}(h(v^i, t^i)),\ t^i\left(x_{RGB}^i \cdot s^i\right)\right) \qquad (2)$$

We use Adaptive Moment Estimation (Adam) algorithm [12] to update network parameters W to reach the optimal value. The Adam algorithm can obtain independent self-adaptive learning rate changing at any time as shown in formula (5) to get faster convergence speed and better convergence performance, instead of maintaining a single learning rate in classical training method of Stochastic Gradient Descent (SGD). The way to update network parameters W is shown in formula (3)(4)(6):

$$\hat{m}_t = m_t / \left(1 - \beta_1^t\right) \qquad (3)$$

$$\hat{v}_t = v_t / \left(1 - \beta_2^t\right) \qquad (4)$$

$$\alpha_t = \alpha \cdot \sqrt{1 - \beta_2^t} / \left(1 - \beta_1^t\right) \qquad (5)$$

$$w_t = w_{t-1} - \alpha \cdot \hat{m}_t / \left(\sqrt{\hat{v}_t} + \varepsilon\right) \qquad (6)$$

where m_t and v_t are the biased estimation of first and second moment, $\hat{m}_t$ and $\hat{v}_t$ are the error correction estimation of first and second moment, β_1 and β_2 are the exponential decay rate, α is the initial learning rate, and ε is the regularization parameter.

We set β_1=0.9 , β_2=0.999 and ε=10^{-6} according to the empirical value. We start with a learning rate α=0.0005 , then reduce it by a half after every 100000 iterations. After 1000000 iterations the training process stopped.

According the input dimensionality of each layer, we use Gaussian noise distribution to initialized the network parameters W, making the variance of the parameters of each layer consistent, as suggested by Kaiming He et al. [13]

III. EXPERIMENT RESULTS AND ANALYSIS

In order to test the proposed approach, we perform a number of experiments to verify the learning and the generative ability of GDDNE and finally determine the type of data augmentation and appropriate corresponding quantitative relationship between the limited key realistic training samples and the diversified viewpoint images generated.

TABLE I. MEASURE OF SIMILARITY BETWEEN IMAGES GENERATIED AND REALISTIC IMAGES OF TEN SPECIFIED INSTANCE OBJECTS

Similarity measurement	Object category									
	Car model	***CD case***	***Deter-gent***	***Extinguis her***	***Glasses box***	***Oscillo-graph***	***Pill case***	***Storage box***	***Tea caddy***	***Vacuum***
HS	97.319%	97.659%	98.360%	97.707%	97.908%	97.008%	98.449%	96.768%	98.391%	98.639%
RMSE	0.00929	0.01104	0.00995	0.00396	0.00313	0.00657	0.00547	0.00456	0.00482	0.00497
CC	0.93479	0.98787	0.98838	0.95334	0.95057	0.96949	0.98026	0.98899	0.96957	0.99480

A. Learning ability of GDDNE

We select 180 viewpoints of each instance object as the key realistic training samples: 30 rotation angles in plane (interval 12 degrees) and 6 angles of depression (0, 15, 30, 45, 60, 75 degree). 10 specified instance objects as shown in Fig.4: car model, CD case, detergent bottle, fire extinguisher, glasses box, oscillograph, pill case, storage box, tea caddy and vacuum. They are common in daily life but are obviously different in appearance features. For each instance object, we train the GDDNE to generate 180 different viewpoint images that are completely consistent with the key realistic training images. We use three types of image similarity measurements to estimate the quality of images generated compared with the realistic images, and the results are shown in TABLE I. The three types of image similarity measurements are the Histogram Similarity (HS), the Root-mean-square error (RMSE) per pixel and the Pearson Correlation coefficient (CC), according to formula (7)-(9) separately:

$$HS(G,S)=\frac{1}{n}\sum_{i=1}^{n}\left(1-\frac{\left|v_{gi}-v_{ti}\right|}{Max(v_{gi},v_{ti})}\right) \tag{7}$$

$$RMSE(G,T)=\sqrt{\frac{1}{n}\sum_{1}^{n}(g_i-t_i)^2} \tag{8}$$

$$CC(G,T)=\frac{\sum_{i=1}^{n}\left(t_i-\bar{t}\right)\left(g_i-\bar{g}\right)}{\sqrt{\sum_{i=1}^{n}\left(t_i-\bar{t}\right)^2}\sqrt{\sum_{i=1}^{n}\left(g_i-\bar{g}\right)^2}} \tag{9}$$

where n is the number of image pixels, v_{gi} and v_{ti} are the number of pixels corresponding to each gray value in the histogram of image generated and true image, g_i and t_i are the per pixel value in the image generated and the realistic image.

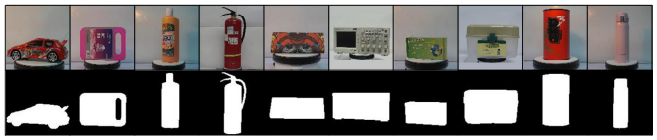

Fig. 4. 10 specified instance objects images (car model, CD case, detergent bottle, fire extinguisher, glasses box, oscillograph, pill case, storage box, tea caddy and vacuum). The first row is the realistic training samples, and the second row is the segmentation masks.

TABLE I. shows that the HS between the images generated and their realistic images of 10 specified instance objects are all above 96.5%, indicat1ing that the network can learn the distribution of image color and shape features well. Besides, the RMSE per pixel between the images generated and their realistic images of 10 specified instance objects are all less than 0.012, which means that GDDNE has a good generation accuracy. Meanwhile, the CC can be used to describe the degree of linear correlation between two distributions, and the value of 10 specified instance objects are all above 0.93, verifying the positive correlation distribution between the images generated and the realistic images is satisfied. To summarize, it can be seen that GDDNE has a good learning ability and can reconstruct the feature distribution of the images in training samples according to the results of three different types of image similarity measurements.

B. Interpolation between rotation angles in plane

We measure the ability of interpolation generation of the GDDNE from two aspects: the accuracy of images generated with different numbers of key realistic training samples; the range of images generated in diversified rotation angles in plane on the premise of guaranteeing the generative precision.

1) Different numbers of key realistic training samples

We select one of the specified instance objects (tea caddy) for representative verification experiments. We use a turntable at a speed of 24.6 round / sec, a camera with a resolution of 1920×1080 and a frame rate of 50fps. For each instance objects, we can finally obtain 1230 realistic images at each depression angle of the camera. Such images as the ground truth realistic images used in the verification of subsequent experiments. We then varied the numbers of the key realistic training samples (180, 90, 48, 24) and train the GDDNE as before to generate 1230 viewpoint images corresponding to the ground truth realistic images under the conditions of 0, 30 and 60 depression angle respectively.

Fig.5 shows some representative examples of rotation angle interpolation generated. For 30 and 15 rotation angles in plane under each 6 depression angle in the key realistic training samples, the effect of angle interpolation generated is satisfactory: the interpolation generative effect is smooth and the feature details are well preserved. However, starting from 8 rotation angles, the GDDNE fails to produce satisfactory interpolation images, which means some pattern features are lost.

In Fig.6, we plot the average RMSE of the generative images under different numbers of key realistic training samples. Obviously, increasing the training samples can improve the generation performance of GDDNE, and when the key training samples increase to 15, GDDNE has been able to

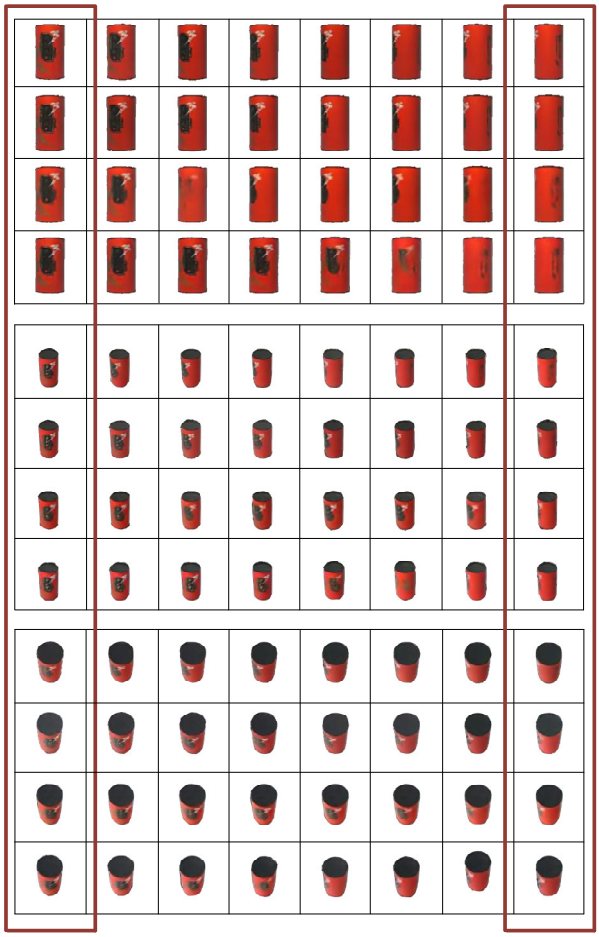

Fig. 5. The leftmost and the rightmost column of images are in realistic training samples, while all intermediate ones are the result of interpolation. Each group of four rows represents depression angle of 0, 30, 60 (top-down), and the number of different training samples is 180, 90, 48, 24 (top-down row in each group).

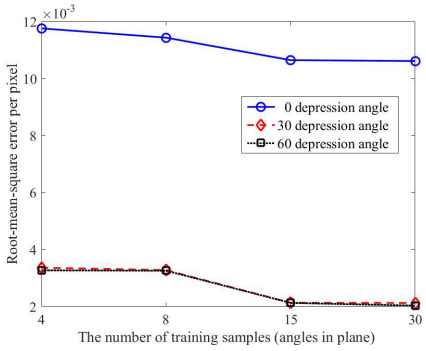

Fig. 6. The average RMSE of the images generated under different numbers of key realistic training samples.

obtain a better result. In addition, the RMSE of the images generated decreases with the increase of depression angle, because the feature informations such as the shape or pattern contained in different depression images are changeable. For tea caddy, the area of the black cover in image will become larger with the increase of depression angle.

2) The range of new viewpoint images generated

This section explores the generative capacity of the GDDNE. We still select the instance object of tea caddy and use the 180 key realistic training samples to train the GDDNE, making the network generate images in 180, 360, 540, 720, 1080 and 1230 rotation angles in plane under the condition of 0, 30 and 60 depression angle respectively.

In Fig.7, we plot the average RMSE of the images generated in different numbers of rotation angles in plane. Obviously, the RMSE of the images generated begin to fluctuate after 900 rotation angles in plane. In addition, a similar situation with Fig.6 is that the RMSE of the generated images will decrease with the increase of depression angle, moreover, the RMSE of average pixel between the different generation numbers of angle images is fluctuating more smoother under the condition of 30 and 60 depression angle.

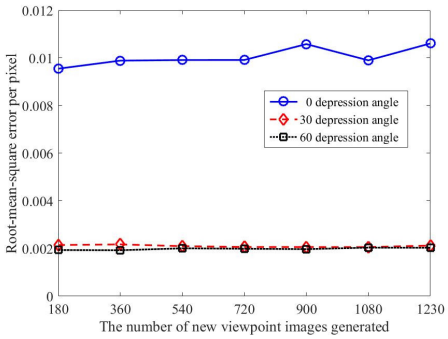

Fig. 7. The average RMSE of the images generated in different numbers of rotation angles in plane.

Overall, taking the tea caddy as representative example, when training with other specified instance objects, the GDDNE can also show a great generation effect with a corresponding quantitative relationship between the key realistic training samples and the viewpoint images generated: more than 90 viewpoints in training samples and within 1200 viewpoint images generated.

C. Data augmentation transformations

Through the experiment, we find that the GDDNE can generate six kinds of transform images – shift horizontally or vertically, zoom, stretching horizontally or vertically, in-plane rotation, brightness and saturation, as well as the superposition effect of different transformations.

In order to guarantee its transformation accuracy, we only select one or two kinds of combination transformation forms, some parts are shown in Fig.8. We find that the quality of the images generated with 1-2 arbitrary transformations is generally great and satisfies the requirements of our specific application task for building a large-scale training data with labels used in a deep network for detecting instance object in unstructured environment.

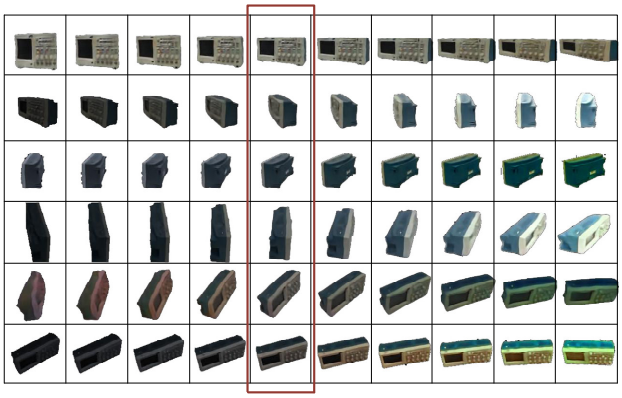

Fig. 8. Each row shows one or two kinds of combination of transformations images: stretching, brightness, saturation, stretching+brightness, in-plane rotation+saturations, brightness+saturation(top-down). The fifth column of images are the realistic non-transformed training samples.

D. Building a large-scale annotated training data

We collect some indoor scene datasets as complex background images to provide unstructured environment information for specified instance objects. For example, the Indoor Scene Recognition Dataset [14], the SUN Dataset [15] and the RGB-D Scenes Dataset [16], a total of 16000 images. Then we select some scene that suitable for our 10 specified instance objects, such as laboratory, classroom, kitchen, conference room and so on.

According to the experimental results above, we finally obtain 1000 diversified viewpoint images for each instance object generated by the GDDNE, with 180 viewpoint images in realistic training samples: 30 rotation angles in plane and 6 angles of depression. At the same time, we arbitrarily superimpose 1-2 data augmentation transformation forms for each images, and the totally order of magnitude is $1000\times\left(1+C_6^1+C_6^2\right)\times 10=220000$. Some parts of annotated images in training data are shown in Fig.9, and the annotation files contain the starting coordinates of boundary frame ($X_{\min}$ and $Y_{\min}$), the sizes of the instance objects (width and height) and the category of the specified instance objects.

Fig. 9. Examples of annotated images in training data (storage box). The top row is images without data augmentation transformations and the second row is with transformations.

IV. CONCLUSIONS

In this article, a Generative Deep Deconvolutional Neural Network (GDDNE) based on the supervised training strategy is proposed for handling the time-consuming issue of artificial acquisition and manual annotation when obtain a large-scale annotated training data for a network to detect the instance object in unstructured environment. The experiment results demonstrate that with 180 key realistic training samples, the GDDNE can automatically generate 1000 diversified viewpoint images and 21 kinds of data transformations for each instance object with great quality. Moreover, an interesting direction for our future research is training the GDDNE to generate more feature change forms, such as different texture or shelter situation.

ACKNOWLEDGMENT

This work was supported by a grant from National Natural Science Foundation of China (61673039).

REFERENCES

[1] Wang R, Liang Y, Xu J W, et al. “Cascading classifier with discriminative multi-features for a specific 3D object real-time detection”. Visual Computer, vol.1, pp.1-16, 2018.

[2] Zeiler M D, Krishnan D, Taylor G W, et al. “Deconvolutional networks”. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp.2528-2535, 2010.

[3] Kulkarni T D, Whitney W F, Kohli P, et al. “Deep convolutional inverse graphics network”, vol.71, pp.2539-2547, 2015.

[4] Dosovitskiy A, Springenberg J T, Brox T. “Learning to generate chairs with convolutional neural networks”. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp.1538-1546, 2015.

[5] [12] Dosovitskiy A, Springenberg J, Tatarchenko M, Brox T. “Learning to generate chairs, tables and cars with convolutional networks”. IEEE Transactions on Pattern Analysis & Machine Intelligence, vol.39, pp.692, 2017.

[6] Radford A, Metz L, Chintala S. “Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks”. Computer Science, 2015.

[7] Stauffer C. “Adaptive background mixture model for real-time tracking”. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol.2, pp.2246, 1999.

[8] Yamato J, Ohya J, Ishii K. “Recognizing human action in time-sequential images using hidden markov model”. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol.J76-D-II, pp.379-385, 1992.

[9] Sutskever I, Hinton G, Taylor G. W. “The recurrent temporal restricted boltzmann machine”. International Conference on Neural Information Processing Systems, vol.20, pp.1601-1608, 2008.

[10] Salakhutdinov R, Hinton G. “Deep boltzmann machines”. Journal of Machine Learning Research, vol.5, pp.1967-2006, 2009.

[11] Jia, Yangqing, Shelhamer, et al. “Caffe: Convolutional Architecture for Fast Feature Embedding”. Proceedings of the 22nd ACM international conference on Multimedia, pp.675-678, 2014.

[12] Kingma D P, Ba J. “Adam: A Method for Stochastic Optimization”. Computer Science, 2014.

[13] He K, Zhang X, Ren S, et al. “Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification”. *2015 IEEE International Conference on Computer Vision*, pp.1026-1034, 2015.

[14] Quattoni A, Torralba A. “Recognizing indoor scenes”. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp.413-420, 2009.

[15] Xiao J, Ehinger K A, Hays J, Torralba A, Oliva A. “Sun database: exploring a large collection of scene categories”. International Journal of Computer Vision, vol.119, pp.3-22, 2016.

[16] Lai K, Bo L, Ren X, Fox D. “Detection-based object labeling in 3D scenes”. *IEEE International Conference on Robotics and Automation*, vol.162, pp.1330-1337, 2012.