Research on Spatial Transformation in Image Based on Deep Learning

Peng Gao[a,] *, Qingxuan Jia[b]

Laboratory of Space Robot, Beijing University of Posts and Telecommunications, No.10 Xitucheng Road, Haidian District, Beijing, China;

[a]gaopeng19921013@gmail.com, [b]qingxuan@bupt.edu.cn

Keywords: image space, spatial transformation, semantic information, image synthesis.

Abstract: In the field of computer graphics, synthesizing a new view of 3D objects in images from a single perspective image is an important problem. A part of the object is unobservable, since 3D objects mapping to image space will result in partial occlusion or self-occlusion of objects. The synthesis needs to infer spatial structure and posture of the object. The uncertainty due to occlusion is a problem in the synthesis. In this paper, the problem is solved by establishing a convolutional neural network (CNN), which uses images including multiple chairs as dataset. First of all, we study related networks to propose a novel multi-parallel and multi-level encoding-decoding network, which implements the transformation from a single perspective image and angle semantic information to a new perspective synthetic image in an end-to-end way. Secondly, the network is trained by establishing a dataset. Finally, it is proved the neural network performs better edge smoothing effect and higher precision in image synthesis than state-of-the-art networks.

1. Introduction

With the rapid development of computer vision technology, the digital image manipulation and processing technology gradually integrates into the daily life. Synthesizing lifelike images has been put forward for a long time by the 3D model and environment variables, which is usually referred to rendering. Meanwhile, recent advances in visual algorithms have enabled computers to understand objects in images[1, 2, 3, 4]. CNN usually use weight sharing models to reduce the number of parameters, which makes spatial transformation perform spatial invariance in a limited scope. However, it causes network not to infer precise spatial relationships, which leads to a long-standing problem in visual field that applying spatial transformation on objects in images is difficult.

Synthetic images can be acquired by image editors[5]. However, spatial information of 3D objects in images can hardly be obtained by traditional image technology, so that it is difficult to synthesize images in which objects have been spatially transformed. Using neural networks to obtain spatial information of 3D objects in images and reconstruct it will improve the information reconstruction, since neural networks perform a good effect on highly nonlinear features. In addition, the authenticity and intelligence in image synthesis is improved. Experiments[6] show reaction time of people judging whether the objects are the same is linearly related to the angle of direction between two objects, which suggests people in judging classes of objects conduct a psychological transformation. Although many progresses of object recognition are inspired by biology, most neural networks rarely explicitly explain psychological transformation. On the contrary, people are focused on invariance of spatial transformation.

Based on the above theory, a feature fusion network is proposed in this paper, which is multi-parallel and multi-level. The network integrates angle semantic information with spatial information encoded by objects in images and obtains synthetic images by deconvolution.

2. Related Work

To solve the problem of spatial transformation driven images synthesis, many related theoretical

researches have been put forward by scholars. The solutions proposed in these studies basically fell into three categories. The first solution was to realize image synthesis through multiple variants of encoding-decoding structure. The second solution was to realize spatial transformation through the capsule network focusing on spatial features. The last solution used methods of traditional feature learning to realize image synthesis.

2.1 Encoding-decoding Network

Jimei Yang proposed a network of recurrent transformation based on weakly supervised learning[7], aiming at the problem of spatial transformation. The inherent information was lost when 3D objects were mapped to image space. Therefore, Jimei Yang proposed a network that was similar to the end-to-end network in this paper. However, the network adopted LSTM[8]. It transformed image synthesis into non-linear angle traversal. Synthetic images of the fixed angle were generated based on the continuous images serialized.

Meanwhile, Tejas D.Kulkarni proposed an inverse graphics convolutional network[9]. The authors used a hybrid encoding-decoding network to learn a representation that was independent of various transformations. The hybrid encoding-decoding network was a structure of directed graph including multi-level convolution and deconvolution operators, which were trained by the SGVB[10] algorithm. The probabilistic distribution was supported as prior information of features, while the network in this paper did not need to support prior conditions. In addition, the network in this paper did not adopt LSTM to improve feature fusion, but designed multi-parallel and multi-level structure to enhance the reuse and fusion of features.

2.2 Capsule Network

Compared with CNN, the capsule network[11] proposed by G.E.Hinton considered that CNN had problems in obtaining data distribution. The capsule network could process complex internal calculation and encapsulate results into a small vector containing features. Each capsule learned to recognize an implicit definition of visual entity under limited perspective conditions and it would output the probability that the entity exists in its finite domain. Although the network in this paper also extracted spatial information to realize spatial transformation in images, the methods of feature fusion adopted in the network was fundamentally different from it.

2.3 Traditional Machine Learning

Weiguang Ding proposed an algorithm of realizing psychological transformation by optimizing transformation distance[12]. CCA[13] were used to simulate relationship between samples. The idea of recognizing objects by exploring feature space was similar to the idea in this paper. However, exploration of transformation space was achieved by fusing semantic information. In addition, Alexey Dosovitskiy also proposed a network used to synthesize images containing chairs of specified type and pose[14]. The network concatenated multiple semantic information to obtain fusion features. The deconvolution was adopted to obtain synthetic images. Meanwhile, restricted boltzmann machines (RBM) and deep boltzmann machines (DBM) can be used in image synthesis[15].

Compared with above networks, there was no need to translate semantic information into other forms in this paper. At the same time, the network in this paper realized feature fusion by multi-parallel and multi-level structure, rather than the single feature fusion in hidden layers.

3. Multi-parallel and Multi-level Encoding-decoding Network

3.1 Description of Network

The goal of this paper is to transform 3D objects in images through the CNN with specified angle semantic information and output images in which objects have been spatially transformed. At present, deep neural network used for synthesis usually combines the encoding-decoding structure. It extracts features by the encoding structure and reconstructs features by the decoding structure. The

difficulties in this paper are as follows: 1) extraction of spatial features, 2) fusion of spatial features and angle semantic features, 3) spatial transformation based on fusion features. Two versions of neural network are proposed in this paper, as shown in Fig 1 and Fig 2. The image synthesis network combined with angle semantic information v1(model v1) and the image synthesis network combined with angle semantic information v2(model v2).

3.1.1 Extraction of spatial features

It is necessary for the encoding structure to reconstruct spatial information of 3D objects, since 3D objects that have been spatially transformed will represent occluded information in original images. Therefore, VGG16 pre-trained is adopted to extract spatial features of input images[16]. Spatial features of objects are decoded by deconvolution[17], so as to realize spatial transformation. Even if the part of neural network that deals with angle semantic information adopts multi-layer fully connected structure, the structure is not used in the part of neural network that deals with image features. The network is suitable for multi-scale images, since it is the fully convolutional network[18] (FCN) (as shown in Fig 2).

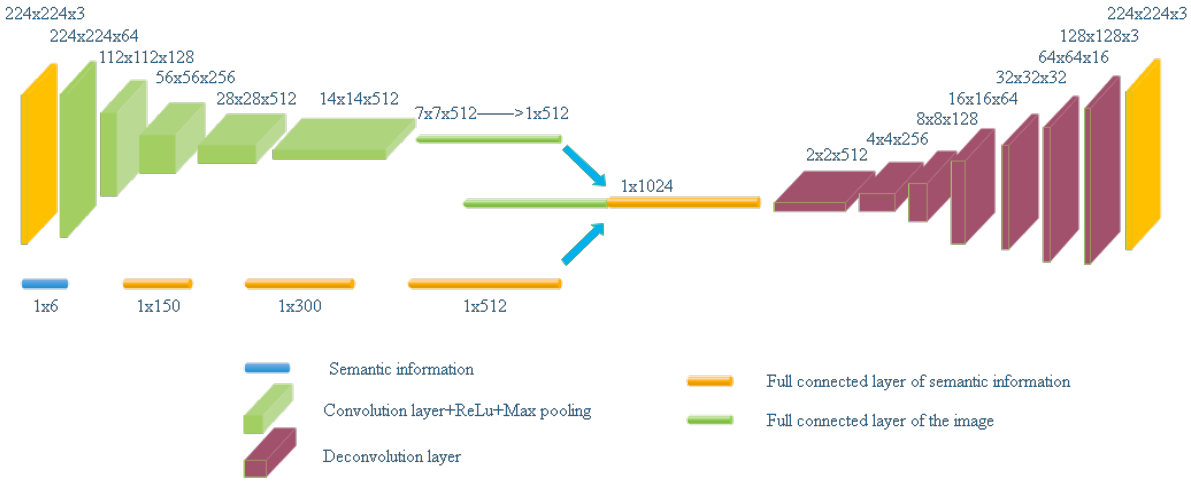

Fig. 1 Image synthesis network combined with angle semantic information v1 (model v1)

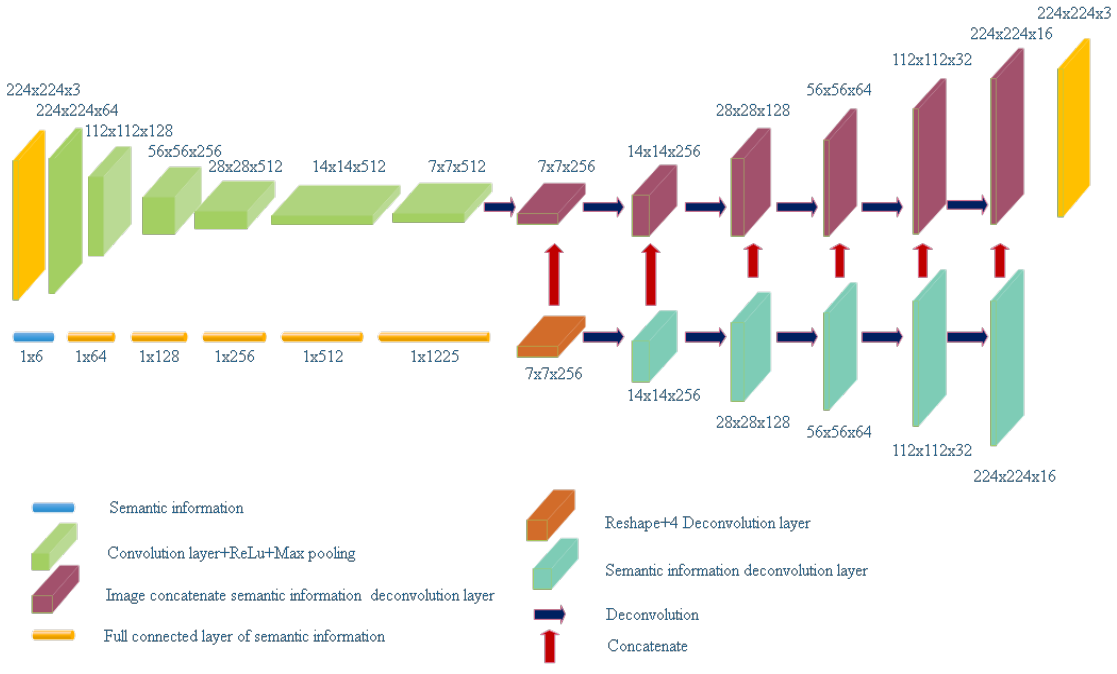

Fig. 2 Image synthesis network combined with angle semantic information v2 (model v2)

3.1.2 The fusion of features

In terms of fusion, spatial features can be transformed to feature vectors based on the network proposed in Alexey Dosovitskiy's paper. The angle semantic information is concatenated by multi-layer fully connected layer. Spatial transformation of 3D objects in images is realized by multi-layer deconvolution (as shown in Fig 1). After analysis, the network has the following problems: 1) the parts occluded in images cannot be recovered. It indicates angle semantic information is not integrated into spatial information. 2) the decoding structure adopts the bilinear interpolation algorithm, which makes images fuzzy. Aiming at the deficiencies of network in Fig 1, we attempt to fuse angle semantic information and image features of different scales in the decoding structure. The network adopts dense structure to combine spatial features with semantic features (as shown in Fig 2), so as to obtain fusion features.

3.1.3 Spatial transformation

The decoding structure further realizes spatial transformation based on feature fusion. Referring to the capsule network, deconvolution is adopted to transform fusion features to synthetic images. The network in Fig 2 applies deconvolution to feature maps formed by angle semantic information and spatial information. Then two feature maps with same number of channels are concatenated on the channel. Deconvolution is carried out on the syncretic feature maps, and so on. Repeated fusion of image features and semantic features can generate image features that have different scales. Such operation can make edges of 3D objects in synthetic images smooth and meticulous. Compared with the network tried in Fig 1, the network in Fig 2 under small angle can also generate smooth synthetic images (as shown in Fig 3, GT is real images, Pre is synthetic images, the angle is between 5° and 85°).

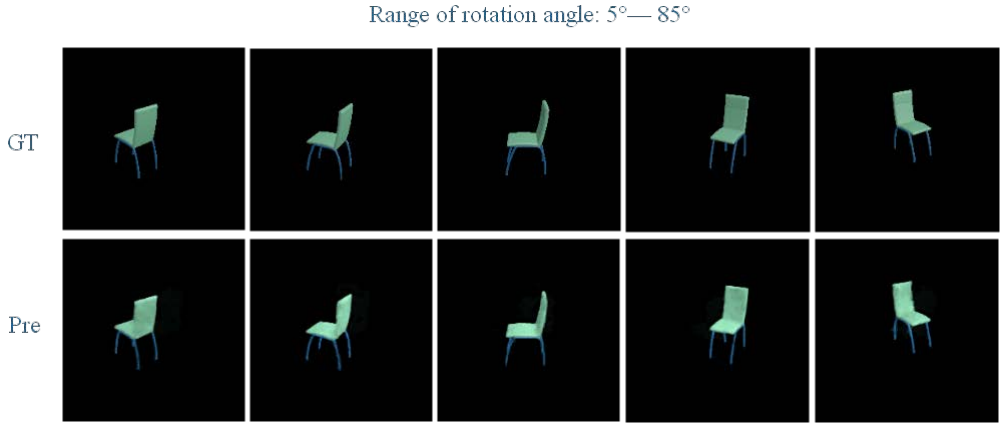

Fig. 3 Testing results based on the network

3.2 Structural Details of Network

At present, the networks generally used to generate images include Generative Adversarial Nets[19] (GAN), Auto-Encoder (AE), Deep Convolution Generative Adversarial Nets[20] (DC-GAN), Information Generative Adversarial Nets[21] (Info-GAN), Variational Auto-Encoder[22] (VAE) and so on. As shown in Fig 2, based on auto-encoder structure, the network is established by a encoding-decoding network according to angle semantic information.

3.2.1 Image encoding-decoding network

The encoding structure adopts VGG16 pre-trained on ImageNet and takes the feature map whose shape is $\left[batch\_size \times 7 \times 7 \times 512\right]$ as output. The decoding structure adopts deconvolution directly. The fully connected layer is not used as the transition layer between encoding and decoding, since images of different scales can be inputted without affecting quality of synthetic images. Therefore, the part of network that processes images is the fully convolutional structure.

3.2.2 Semantic encoding-decoding network

The network uses five fully connected layers to amplify channels of the semantic vector whose shape is $[batch\_size \times 1 \times 6]$ and gets the semantic vector of $[batch\_size \times 1 \times 1225]$. The process is realized by using multi-layer fully connected layers instead of one-layer fully connected layer, since expression bottleneck is caused by extensive expansion of rotation angle, which affects features representation.

3.2.3 Multi-parallel and multi-level decoding structure

The decoding structure adjusts features vectors to four dimensions, whose shape is $[batch\_size \times 7 \times 7 \times 25]$. Four deconvolutional layers are adopted to obtain feature maps of $[batch\_size \times 7 \times 7 \times 256]$ (kernel is 3×3, stride is 1×1). The feature maps are concatenated with the image feature maps (shape is $[batch\_size \times 7 \times 7 \times 256]$) produced by deconvolution on the channel to obtain feature maps of $[batch\_size \times 7 \times 7 \times 512]$. The feature maps are taken as image feature maps for next deconvolutional layer. The feature maps of $[batch\_size \times 14 \times 14 \times 256]$ are obtained and concatenated with semantic feature maps of the same shape, and so on. The deconvolution in decoder consists of 6 layers, in which the shape of feature maps outputted from each layer is similar to the shape of feature maps outputted from each pooling layer of VGG16. However, the network is not completely symmetric. It only ensures that output of the final convolutional layer is consistent with size of the original image and avoids correcting images size by interpolation or filling, so as to facilitate the calculation of subsequent error function and error backpropagation. The multi-parallel and multi-level structure in the network can fully reuse the features, which reduces the number of network layers. Thus, it can accelerate convergence of the network during training.

4. Training Process of Network

4.1 Dataset

This paper employs chairs model of six different types to generate a dataset. At first, based on the chairs of six different types, Blender software is used to synthesize 60,000 images of random angle as training sets (10,000 for each type). At the same time, 60,000 pieces of angle semantic information are generated randomly (10,000 for each type). Finally, based on the images of training sets and corresponding semantic information, Blender is used to synthesize 60000 images that have been spatially transformed as tags for loss calculation. The angle semantic information is a vector of 1×6 (translation component in 3 directions, rotation component in 3 directions), which controls rotation angle of objects in images around the z axis that is z axis at the origin of objects spatial coordinate system. If we only realize translation, it will not have the problem of information occlusion and space transformation in above analysis. Based on the method of establishing training sets, testing sets including 3000 images (500 for each type) and corresponding angle semantic information (500 for each type) are also established to evaluate the network and compare with other networks.

During establishing the training sets, this paper limits rotation angle in (85°, 360°). However, the rotation angle is limited in (1°, 85°) for the testing sets. The network trained is easy to overfit within the scope of 360 ° due to 10000 groups of samples. Therefore, the rotation angle of testing sets and training sets are separated. The mean square error is used as the loss function for backpropagation during training, as shown in formula (1).

$$\frac{1}{n}\sum_{t=1}^{n}\frac{1}{N}\sum_{i=1}^{N}\left(y_i^t - \hat{y}_i^t \right) \tag{1}$$

where, y_i^t is the pixel value of synthetic images rendered by Blender software under specified

angles, which can be used as the output label of original images and semantic information. Meanwhile, $\hat{y}_i^t$ is the pixel value of images synthesized by the network in Fig 2, which can be used as the output prediction based on original images and semantic information. In addition, n is batch size of the network training, N is the number of pixels.

Based on the training sets established, the batch size is set to 10. After 100 iterations, the loss attenuation line chart shown in Fig 4 can be obtained. From Fig 4, it can be seen that loss has been reduced to 1.8×10^{-4} after 50 iterations and approximated convergence. The network sets the learning rate to 0.001 and conducts training in the way of half attenuation every 50 iterations[23]. In this way, it can accelerate convergence in early training stage and guarantee a good convergence effect in later stage. Adam optimization function is selected during training so as to accelerate convergence of the network[24], which is suitable for optimization of generation model.

The channels of feature maps outputted from each layer in decoding structure are adjusted. It is found that channels of feature maps outputted by front layers of the network are better to be slightly large through experiments, since detailed features are reflected in front layers. The large channels can make the network extract more spatial features in psychological transformation.

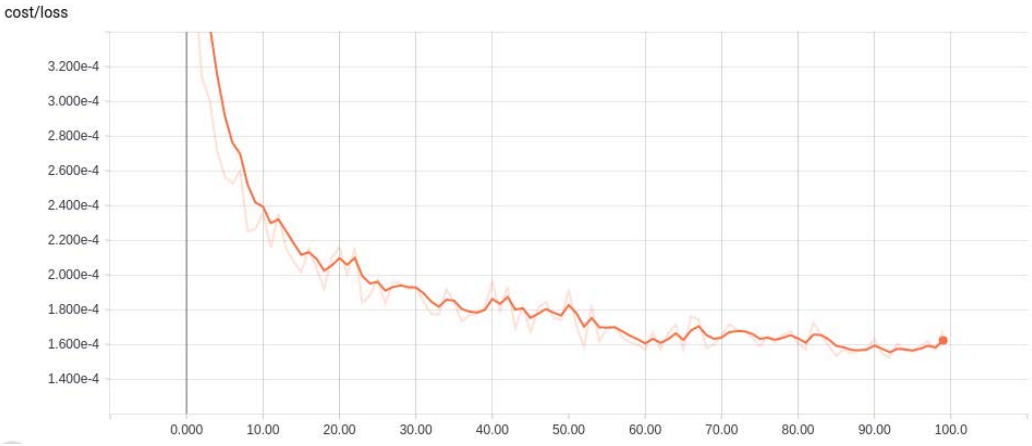

Fig. 4 Loss attenuation line chart

4.2 Testing

The testing is carried out through the network trained based on testing sets including 3000 groups images. From the experimental results (as shown in Fig 5), it can be seen that the network has learned ability to obtain spatial information of 3D objects in images and understand angle semantic information. Furthermore, the network can also apply spatial transformation to 3D objects in images with angle semantic information. The experimental results show that no matter the rotation angle is between 5° and 85° or between 1° and 5°, the network can generate clear images which meet angle semantic information.

Weight parameters of the network are about 50Mb, which are loaded to achieve the goal of synthesizing images for all kinds of chairs. According to the results in Fig 5, the network can recover information of texture and color after spatial transformation of various chairs. Meanwhile, it reconstructs occluded parts of original images. Moreover, the results in Fig 3 and Fig 5 show that the network can make objects edges in images smooth and meticulous. Finally, the positioning accuracy of synthetic images is relatively high. The network will generate slightly blurred synthetic images only for chairs with dense structure.

The comparison between Fig 5 shows that although rotation angle between 1° and 5° has higher requirements for synthesizing images than rotation angle between 5° and 85°, the synthetic images corresponding to rotation angle between 1° and 5° performs relatively high precision. Therefore, the network can also perform perfect effects for small rotation angle.

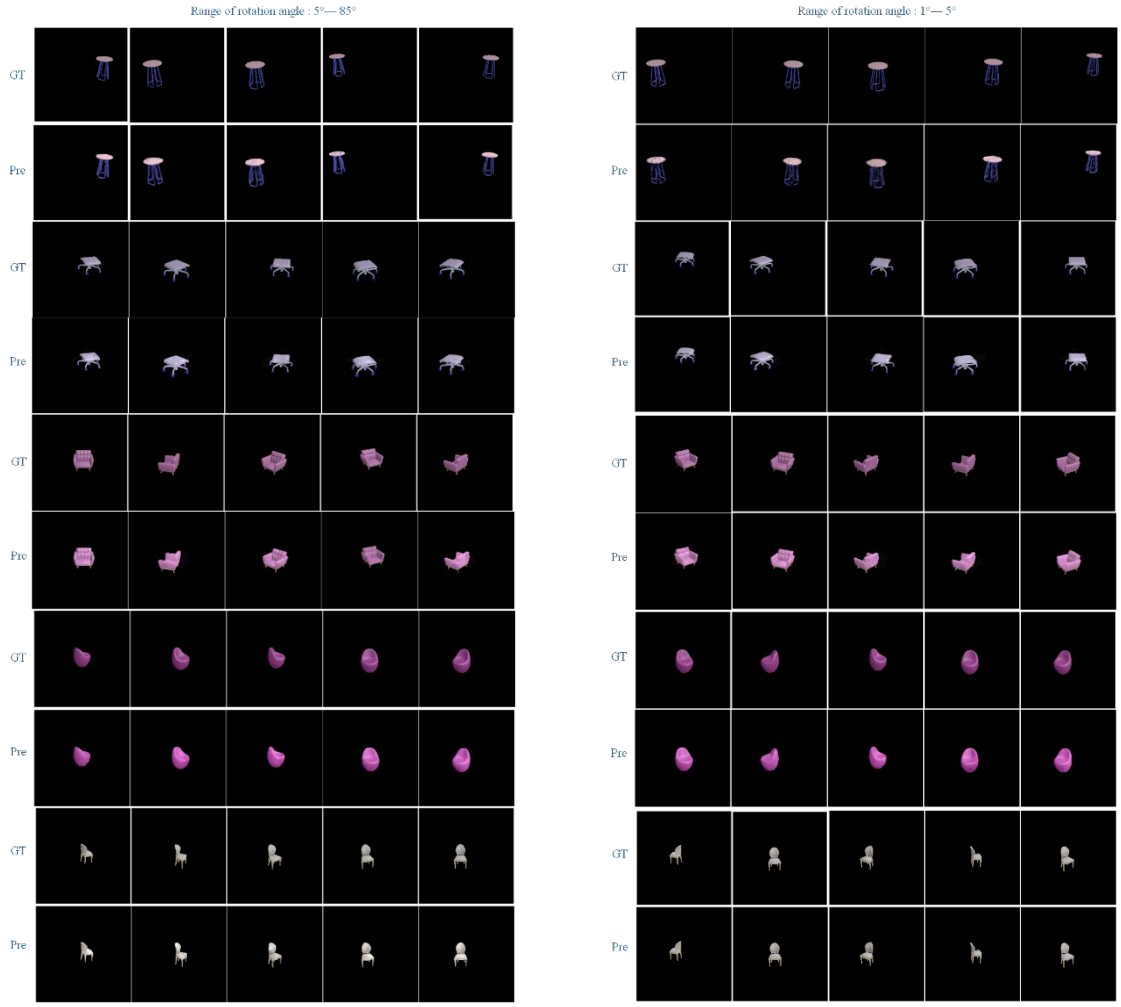

Fig. 5 Experimental results of different angles

5. Experimental Comparison and Analysis

5.1 Experimental Comparison

5.1.1 Qualitative comparison

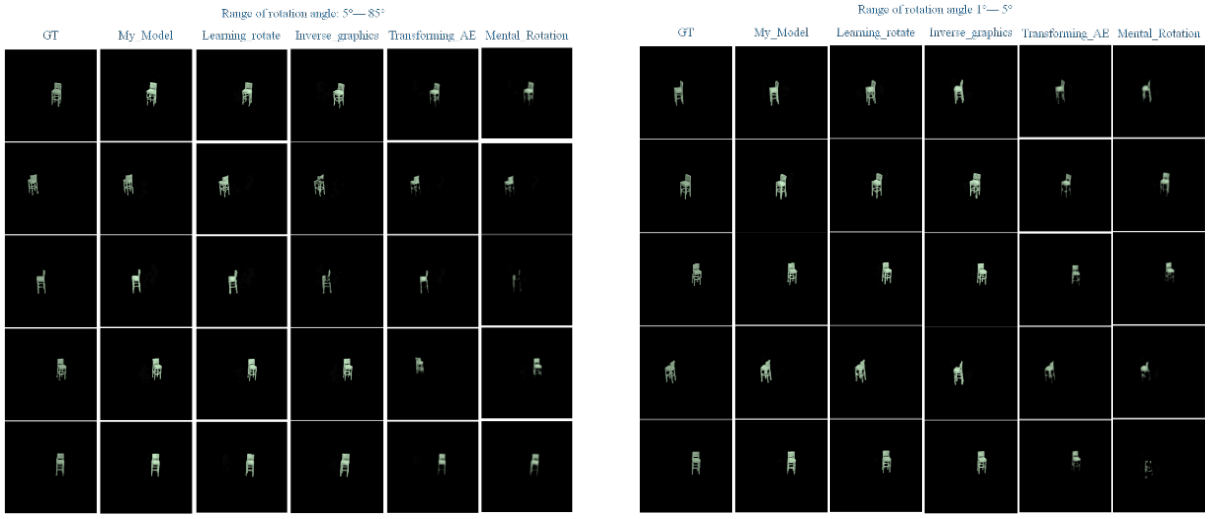

Fig. 6 Prediction comparison of different angles

The effects of networks proposed in related researches and the network in this paper on

synthesizing images are compared based on one type of chairs in testing sets. The result is shown in Fig 6, it compares the network in this paper with other four networks in the rotation angle between 1° and 5° and between 5° and 85°. It is necessary to transform the angle semantic information, since each network requires different formats to satisfy semantic conditions. For the network proposed by Jimei Yang (Learning_rotate in Fig 6), semantic information needs to be converted to the format of act units in the paper. The information of act units is added to hidden layers of the network and synthetic images of fixed angle are generated by RNN. For the capsule network (Transforming_AE in Fig 6), the mental rotation transformation network (Mental_Rotation in Fig 6), and the inverse graphics network (Inverse_graphics in Fig 6), semantic information needs to be converted to variable forms corresponding to hidden layers.

5.1.2 Quantitative comparison

This paper uses the block histogram matching algorithm to calculate similarity of images, since the histogram matching algorithm cannot fully represent the position information of images. Firstly, the image is divided into 16 blocks. Then, the histogram matching value of each block is calculated according to formula (2). Finally, the average is used to obtain the final similarity.

$$Sim(G,S)=\frac{1}{N}\sum_{i=1}^{N}\left(1-\frac{|g_i-s_i|}{Max(g_i,s_i)}\right) \tag{2}$$

where, G is the set of real images and S is the set of synthetic images.

Combined with the block histogram matching algorithm, this paper quantitatively evaluates the similarity between synthetic images and real images of each network. As shown in table 1 and table 2, similarity between synthetic images and real images in different ranges of angles is obtained by the algorithm, in which the great value represents the high similarity. Model v1 is version 1 of the network in this paper and model v2 is version 2 of the network in this paper.

Table 1 Similarity of synthetic images and real images whose angle between 1° and 5°

model v1	model v2	Learning_rotate	Inverse_graphics	Transforming_AE	Mental_Rotation
0.4356	0.8818	0.8716	0.8796	0.8581	0.8482
0.4231	0.9253	0.9191	0.9045	0.8611	0.8356
0.5231	0.9246	0.9337	0.9199	0.8345	0.8243
0.4361	0.9279	0.9167	0.9087	0.8456	0.8325
0.5362	0.9168	0.9081	0.8983	0.8320	0.8269
0.4625	0.9156	0.9272	0.9146	0.8424	0.8317
0.3856	0.8923	0.9023	0.9015	0.8638	0.8409
0.4361	0.9155	0.8956	0.8798	0.8206	0.8112
0.4289	0.9026	0.9031	0.8876	0.8353	0.8324
0.4625	0.9156	0.9042	0.9085	0.8482	0.8256

Table 2 Similarity of synthetic images and real images whose angle between 5° and 85°

model v1	model v2	Learning_rotate	Inverse_graphics	Transforming_AE	Mental_Rotation
0.4623	0.9285	0.9244	0.8979	0.9011	0.8973
0.5635	0.9483	0.9371	0.9340	0.9043	0.9130
0.5472	0.9335	0.9257	0.9334	0.9142	0.9048
0.3561	0.9294	0.9151	0.9154	0.8965	0.8992
0.3281	0.9350	0.9430	0.9117	0.8898	0.9086
0.6532	0.9324	0.9141	0.9268	0.8993	0.8872
0.4328	0.9353	0.8945	0.8802	0.8890	0.8755
0.5192	0.9345	0.9315	0.9251	0.9055	0.8821
0.5364	0.9341	0.9456	0.9102	0.9045	0.8995
0.4351	0.9338	0.9231	0.9214	0.8921	0.8826

5.2 Results Analysis

At first, the network in this paper can obtain angle features from original semantic information and integrate angle features into spatial features without transformation and constraint. Secondly, the network can synthesize images of high similarity. Compared with the capsule network and the mental rotation transformation network, the network performs high similarity.

According to table 1 and table 2, the quality of synthetic images from model v1 is poor and the quality of synthetic images from model v2 is significantly improved compared with other models. Relative to the recurrent transformation network and the inverse graphics network, the quality of synthetic images from the capsule network and the mental rotation transformation network is lower. Similarity corresponding to rotation angle between 5° and 85° are generally higher than rotation angle between 1° and 5°, since spatial transformation of small angles are more difficult than large angles. Finally, the network proposed by Jimei Yang can only output synthetic images of fixed angles. Meanwhile, the network generated by Alexey Dosovitskiy can only generate images by interpolation as well, which are worse than images generated directly from networks. However, the model v2 can carry out spatial transformation of 3D objects in images in accordance with arbitrary angles and its precision can reach 1°.

6. Conclusions

In this paper, a multi-parallel and multi-level encoding-decoding network is proposed to realize spatial transformation of 3D objects in images combined with angle semantic information. Based on structural analysis of the network and the experimental results, this paper performs the following three contributions: First of all, the network in this paper can reconstruct 3D objects in images from image space and recover the spatial information and texture information, so that it can recover the occluded information in synthetic images. Secondly, the network can fuse image features and semantic features of different scales by a dense structure that is multi-parallel and multi-level. Finally, the network can implement spatial transformation of 3D objects in image space by multi-layer feature fusion.

Acknowledgments

This work was financially supported by National Science Foundation of China (NSFC No.61573066). Peng Gao is the corresponding author of this paper.

References

[1] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]// International Conference on Neural Information Processing Systems. Curran Associates Inc. 2012:1097-1105.

[2] Girshick R. Fast R-CNN[J]. Computer Science, 2015.

[3] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2015:3431-3440.

[4] Vinyals O, Toshev A, Bengio S, et al. Show and tell: A neural image caption generator[C]// Computer Vision and Pattern Recognition. IEEE, 2015:3156-3164.

[5] Kholgade N, Simon T, Efros A, et al. 3D object manipulation in a single photograph using stock 3D models [J]. ACM Transactions on Graphics, 2014, 33(4):1-12.

[6] Shepard RN, Metzler J. Mental rotation of three-dimensional objects[J]. Science, 1971, 171(3972):701-703.

[7] Yang J, Reed S, Yang M H, et al. Weakly-supervised dis-entangling with recurrent transformations for 3D view synthesis[C]// International Conference on Neural Information Processing Systems. MIT Press, 2015:1099-1107.

[8] Li X, Wu X. Constructing long short-term memory based deep recurrent neural networks for large vocabulary speech recognition[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2015:4520-4524.

[9] Kulkarni T D, Whitney W F, Kohli P, et al. Deep convolutional inverse graphics network[C]// International Conference on Neural Information Processing Systems. MIT Press, 2015:2539-2547.

[10] Lopez R, Regier J, Yosef N, et al. Information Constraints on Auto-Encoding Variational Bayes [J]. 2018.

[11] Hinton G E, Krizhevsky A, Wang S D. Transforming Auto-Encoders[C]// International Conference on Artificial Neural Networks. Springer, Berlin, Heidelberg, 2011:44-51.

[12] Chen, Kaisheng, Huang, Zhuyang, Dong, Gaoneng, et al. "Mental Rotation" by Optimizing Transforming Distance [J]. Solid State Communications, 2014, 21(1):25–32.

[13] Rasiwasia N, Mahadevan V, Aggarwal G, et al. Cluster Canonical Correlation Analysis [J]. Aistats, 2014.

[14] Dosovitskiy A, Springenberg J T, Brox T. Learning to generate chairs with convolutional neural networks[C]// Computer Vision and Pattern Recognition. IEEE, 2015:1538-1546.

[15] Tan S, Mayrovouniotis M L. Reducing data dimensionality through optimizing neural network inputs [J]. Aiche Journal, 1995, 41(6):1471-1480.

[16] Sainath T N, Kingsbury B, Saon G, et al. Deep Convolutional Neural Networks for Large-scale Speech Tasks [J]. Neural Networks, 2015, 64:39-48.

[17] Zeiler M D, Krishnan D, Taylor G W, et al. Deconvolutional networks[C]// Computer Vision and Pattern Recognition. IEEE, 2010:2528-2535.

[18] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]// IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, 2015:3431-3440.

[19] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[C]// International Conference on Neural Information Processing Systems. MIT Press, 2014:2672-2680.

[20] Radford A, Metz L, Chintala S. Unsupervised Representation Learning with Deep Convolutional Generative Adver-sarial Networks [J]. Computer Science, 2015.

[21] Chen X, Duan Y, Houthooft R, et al. InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets [J]. 2016.

[22] Bodin E, Malik I, Ek C H, et al. Nonparametric Inference for Auto-Encoding Variational Bayes[J]. 2017.

[23] Bengio Y, Collobert R, Weston J. Curriculum learning[C]// International Conference on Machine Learning. ACM, 2009:41-48.

[24] Im D J, Kim C D, Jiang H, et al. Generating images with recurrent adversarial networks[J]. 2016.