Generative ConvNet

with Continuous Latent Factors

Author : **Yifei Xu** ; Mentor : Professor **Ying Nian Wu** ; Supervisor : **Jianwen Xie, Tian Han**

1. Introduction

1.1 Background & Objective

Machine learning is nowadays a fundamental part of Artificial Intelligence. People want the computer to learn the real meaning of the image (or other data) by itself.

The generative model can synthesis new data which is similar but not the same by training on sort of unlabeled data. The meaningfulness of the output represent that the model truly understood the input data.

In this project, the target is to design a generative model which can accomplish both reconstruction and synthesis in order to prove its ability on understanding input data. Many experiments are designed to evaluate the effectiveness of this model.

1.2 Model Design

We use the **unsupervised learning** of a popular top-down generative ConvNet model with latent continuous factors can be accomplished by a learning algorithm that consists of alternatively performing **back-propagation** by **Langevin Dynamic** on both the latent factors and the parameters.

The model is a non-linear generalization of **factor analysis**, where the high-dimensional observed data vector, such as an image, is assumed to be the noisy version of a vector generated by a non-linear transformation of a low-dimensional vector of continuous latent factors.

2. Model Structure

2.1 Factor Analysis

The basic linear factor analysis model is

$$Y = WZ + \epsilon$$

Where Y represent observed data vector, $Z \sim N(0, I_d)$ are latent factors, W is transformation matrix and $\epsilon \sim N(0, \sigma^2 I_d)$ is observational noise.

2.2 ConvNet Loading

Our model generalizes the linear loading WZ to a non-linear loading $f(Z;W)$, where f is a ConvNet.

$$Y = f(Z;W) + \epsilon$$

Specifically, we can write the top-down ConvNet as follows

$$f(Z;W) = f_1(W_1 f_2(W_2 \dots f_n(W_n Z + b_n) + \cdots) + b_1)$$

Where f_l, W_l, b_l represent the element-wise non-linearity, weights, bias terms at layer l.

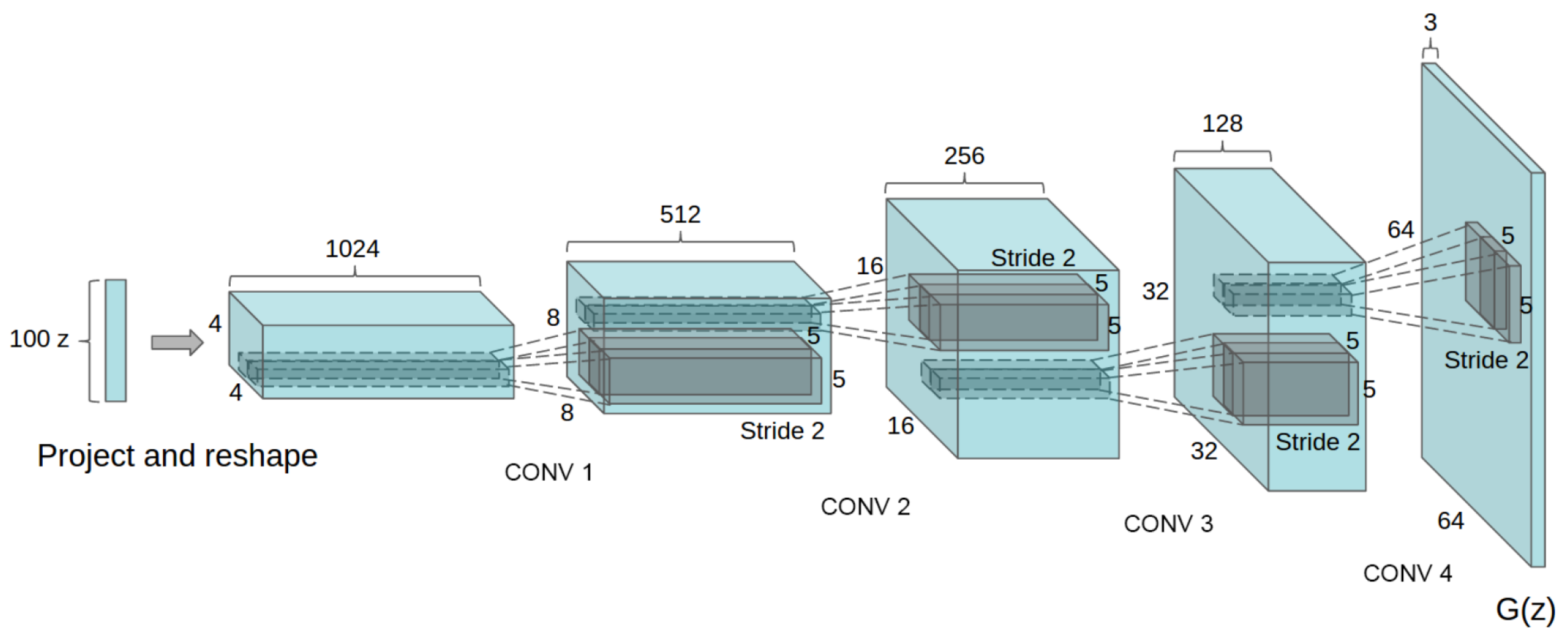

Fig. 1 The structure of ConvNet

3. Alternating back-propagation

3.1 Gradient Descent

If we observe a training set of data vectors $\{Y_i, i = 1,2,\dots,n\}$, then each Y_i has a corresponding Z_i, but all the Y_i share the same ConvNet W. Intuitively, we should infer $\{Z_i\}$ and learn W to minimize the reconstruction error $\sum_{i=1}^{n} \| Y - f(Z_i;W) \|^2$ plus a regularization term that corresponds to the prior on Z. In other word, maximum the log-likehood:

$$L(W, \{Z_i\}) = \sum_{i}^{n} \log p(Y_i, Z_i; W)$$
$$= -\sum_{i}^{n} \left[\frac{\| Y - f(Z_i;W) \|^2}{2\sigma^2} + \frac{\| Z_i \|^2}{2} \right] + constant$$

Whose gradient for W and Z_i is

$$\begin{cases} \dfrac{\partial L}{\partial Z_i} = \dfrac{1}{\sigma^2}\big(Y_i - f(Z_i, W)\big)\dfrac{\partial f}{\partial Z_i} - Z_i \\ \dfrac{\partial L}{\partial W} = \displaystyle\sum_{i=1}^{n} \dfrac{1}{\sigma^2}\big(Y_i - f(Z_i, W)\big)\dfrac{\partial f}{\partial W} \end{cases}$$

Maximum likelihood estimation can be accomplished by the alternating gradient descent that iterates the following two steps, (1) Inferential back-propagation: For each I, run L steps of gradient descent to update $Z_i \leftarrow Z_i + \eta \partial L / \partial Z_i$

(2) Learning back-propagation: Update $W \leftarrow W + \eta \partial L / \partial W$

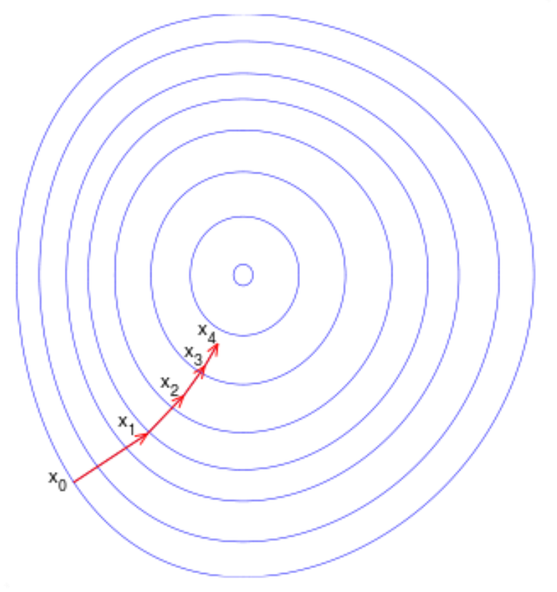

Fig. 2 Gradient Descent in 2D

3.2 Langevin Dynamics

A more rigorous approach to maximum likelihood learning involves maximizing the observed data log-likelihood $L(W) = \sum_{i=1}^{n} \log p(Y_i;W) = \sum_{i=1}^{n} \log \int p(Y_i, Z_i; W) dZ_i$, where the latent factors have been integrated out.

The formula for updating W are the same as gradient descent and updating Z are similar to it, with a noise added.

$$Z_{i+1} = Z_i + \frac{\Delta^2}{2}\left[\frac{1}{\sigma^2}\big(Y - f(Z_i;W)\big)\frac{\partial f}{\partial W} - Z_i\right] + \Delta\epsilon$$

4. Experiment

Our model can be used on any kind of data include sounds, images and videos. The following is the images data experiment.

4.1 Reconstruction and Synthesis

We evaluate our model by reconstruction and synthesis. Theoretically, every $Z \sim N(0, I_d)$ can lead to a meaningful image by train on image. Using Z inferred by an original image is called reconstruction and using random Z is called synthesis.

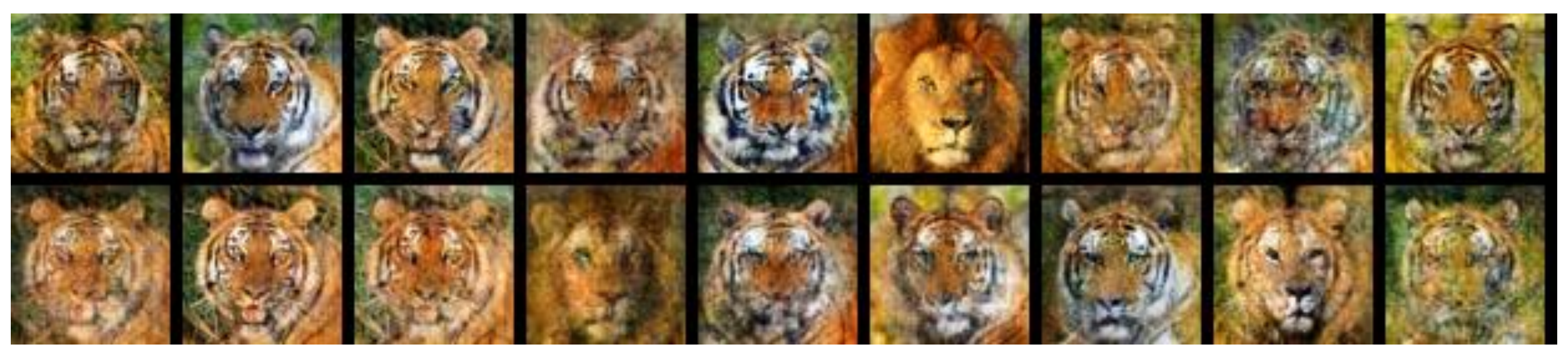

Fig 3. Synthesis Result on on 6 tigers and 5 lions

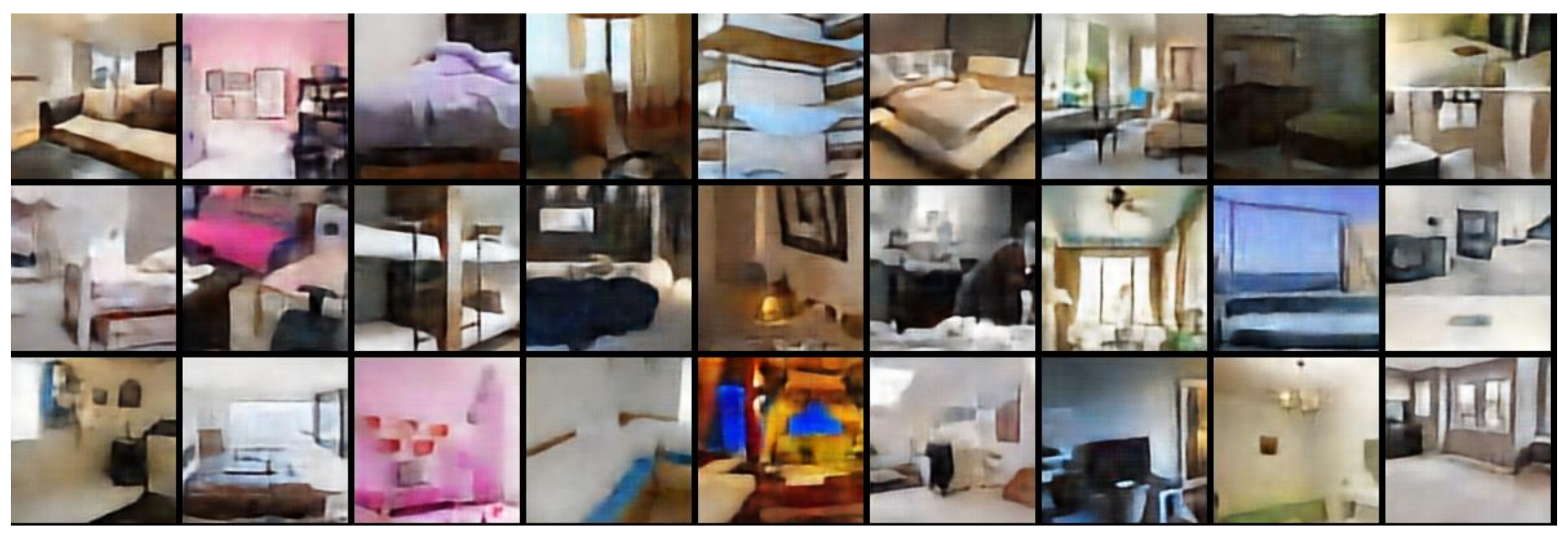

Fig 4. Reconstruction Results on 20000 Bedroom images

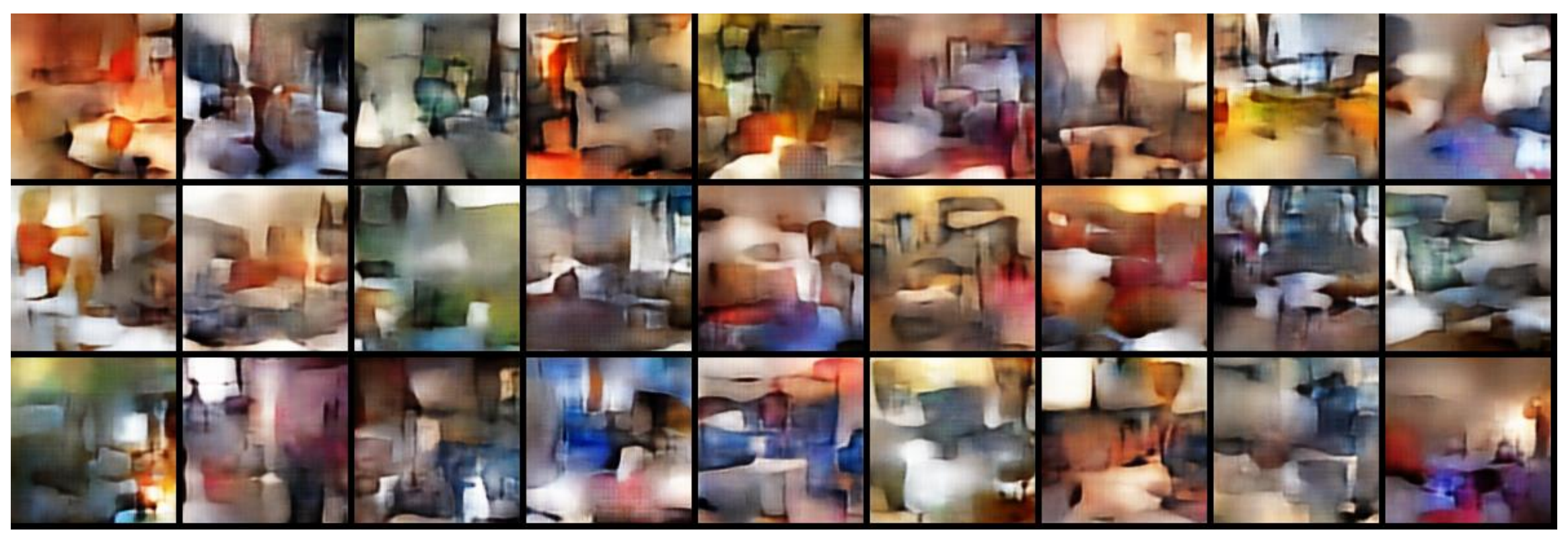

Fig 5. Synthesis Results on 20000 Bedroom images

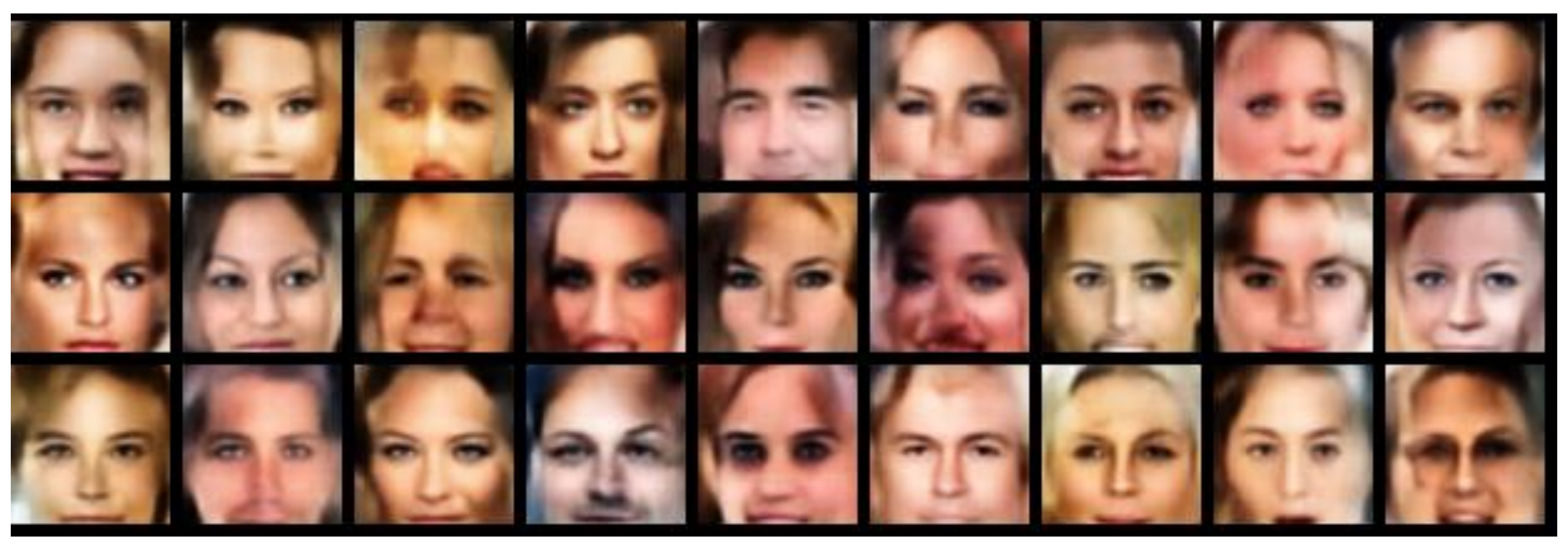

Fig 6. Synthesis Results on 10000 face images

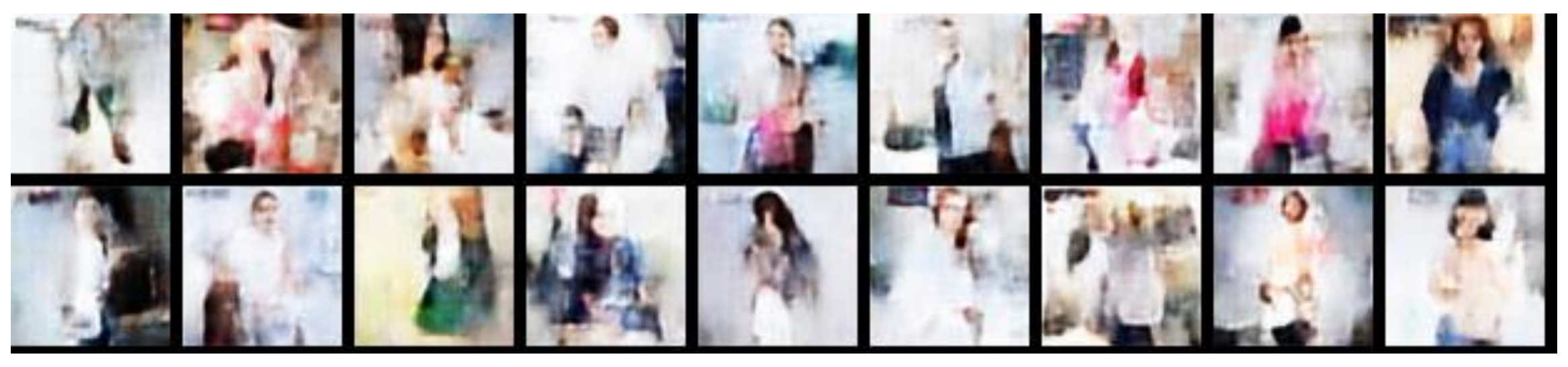

Fig 7. Synthesis Results on 10000 clothes images

4.2 Interpolation

The linear combination of two or more inferred Z (by training image) is called interpolation.

The combination is $Y_i = f(\frac{i}{n} Z_1 + \frac{n-i}{n} Z_2; W)$

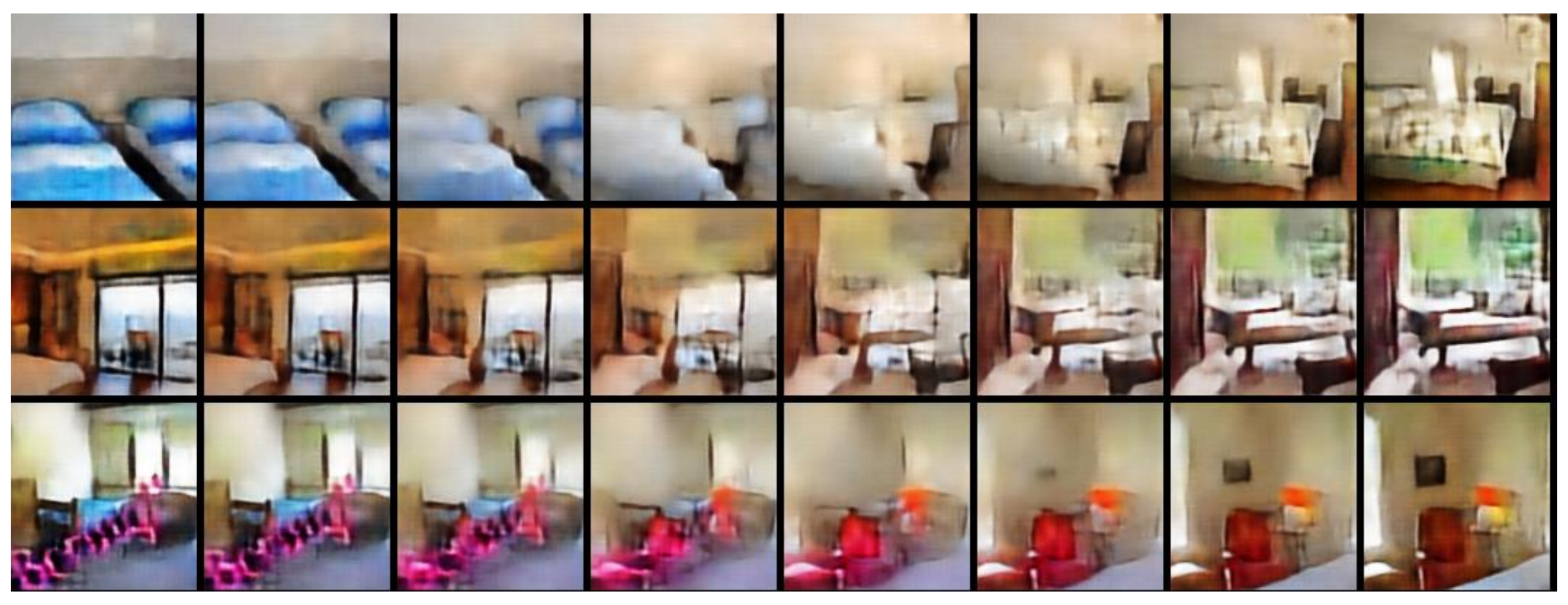

Fig 8. The linear interpolation by 2 images. The left/right most one is the original image.

4.3 Train/test/negative reconstruction

The reconstruction error is defined as

$$\sum_{i=1}^{n} \| Y_i - f(Z_i;W) \|^2$$

In order to evaluate the qualify of the model. We test not only the training data, but also the test data which is the same category as training one but do not involve the training procedure and the negative data which is the different category. The following are 3 model training on cat, face and bedroom. Which are overfitting, good and underfitting.

Trainging on 100 cat images --- Overfitting

Due to small-scale data and large parameters, the training error become slightly small but the test error are as large as the negative error.

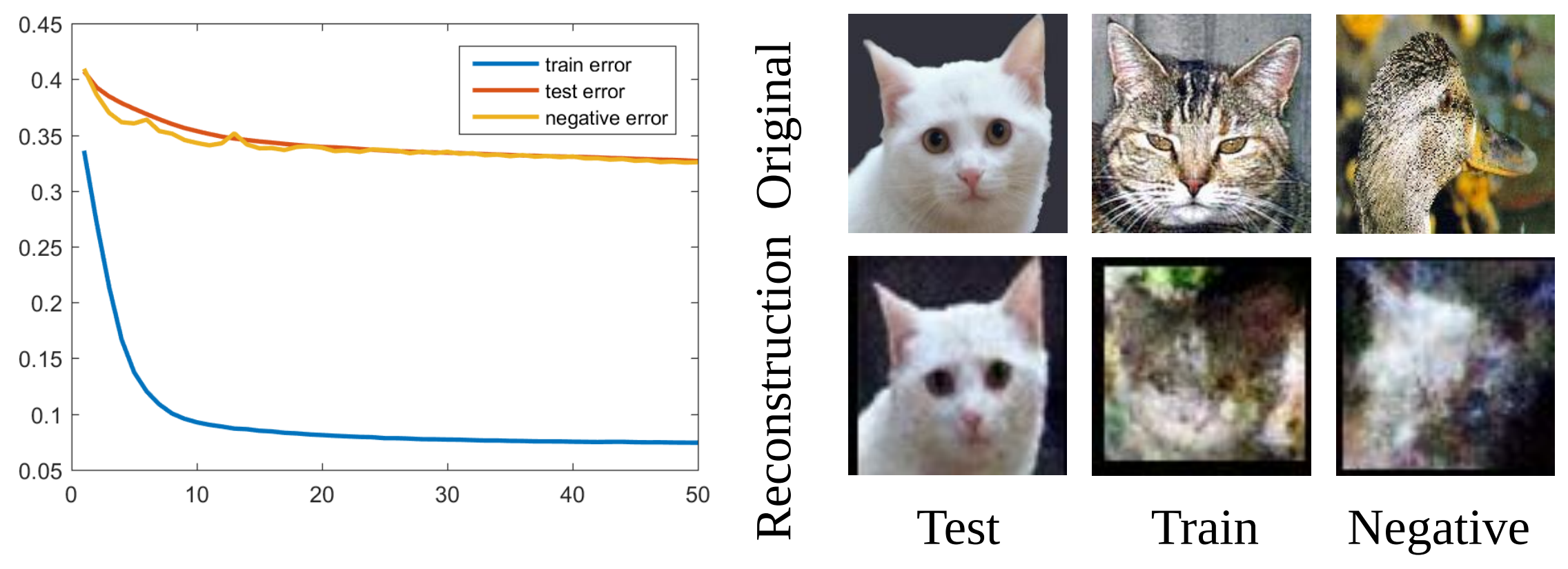

Fig 9. The reconstruction error curve and reconstruction result on cat

Trainging on 10000 face images --- Good result

The result of 10000 face images is good with a low train/test error and a high negative error.

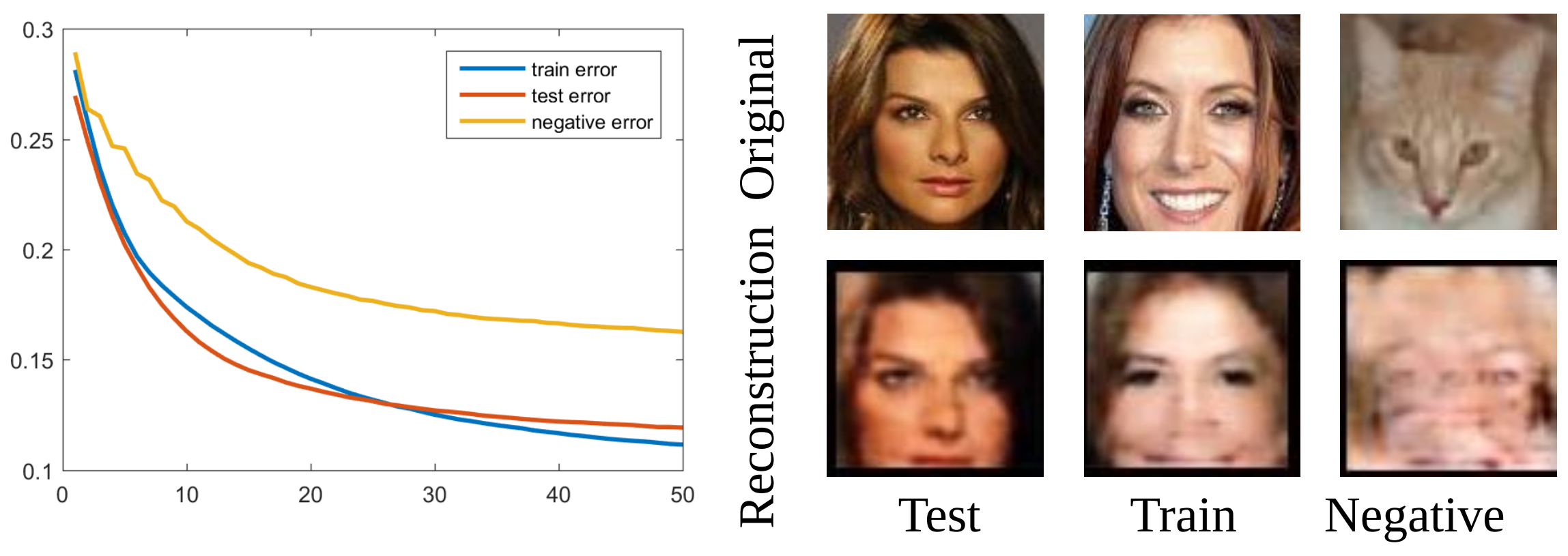

Fig 10. The reconstruction error curve and reconstruction result on face

Trainging on 20000 face images --- Underfitting

The result of 20000 face images is underfitting since train/test/negative error are all similar.

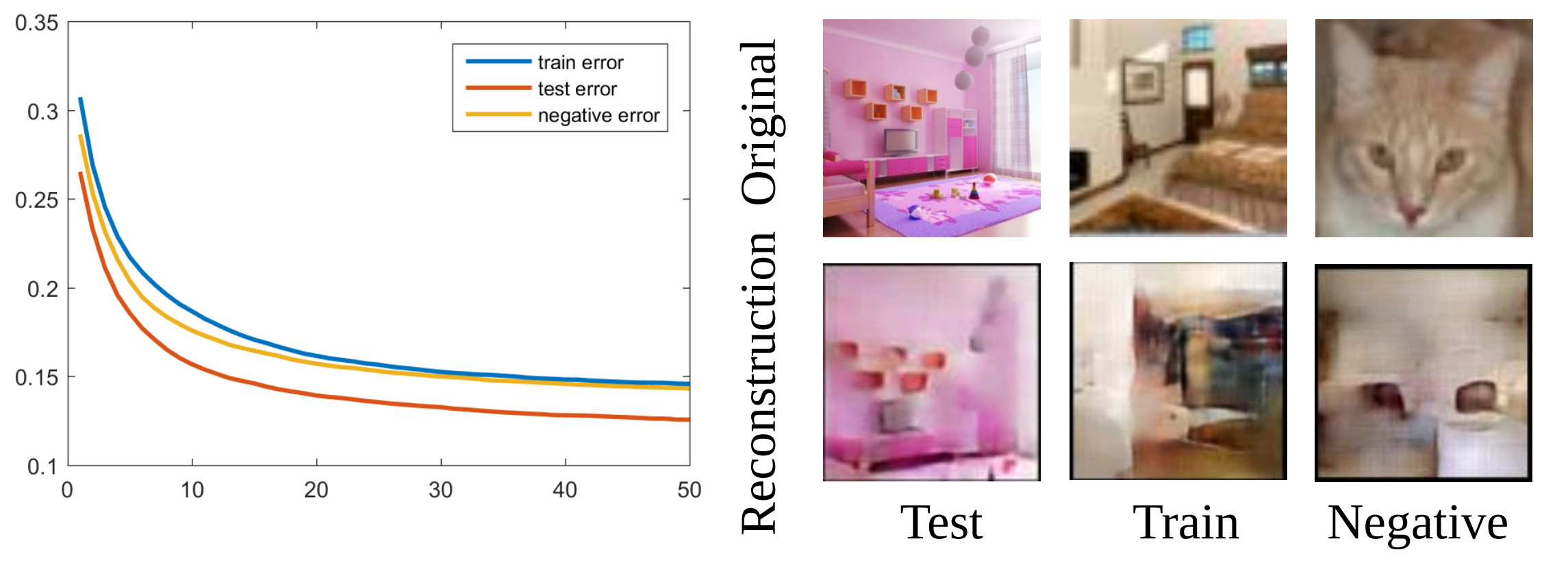

Fig 11. The reconstruction error curve and reconstruction result on bedroom

5. Acknowledgement

Han, Tian, et al. "**Learning Generative ConvNet with Continuous Latent Factors by Alternating Back-Propagation.**" *arXiv preprint arXiv:1606.08571*(2016).

Xie, Jianwen, et al. "**A theory of generative convnet.**" *arXiv preprint arXiv:1602.03264* (2016).

Karpathy, Andrej, et al. "**Large-scale video classification with convolutional neural networks.**" *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition.* 2014.

Radford, Alec, Luke Metz, and Soumith Chintala. "**Unsupervised representation learning with deep convolutional generative adversarial networks.**" " *arXiv preprint arXiv:1511.06434* (2015)

Special thanks to: Yang Lu
Cross-disciplinary Scholars in Science and Technology (CSST)

Visit it for paper, codes and more results:
http://fei22.cn/ABP.html