IA-pix2seq：一个实现简笔画可控生成的深度双向学习方法

臧思聪　涂仕奎　徐雷

(上海交通大学 计算机科学与工程系 上海 200240)

摘　要　构建一个高斯混合模型（GMM）分布的编码空间是一种可辅助实现简笔画可控生成的编码方法。每种特定风格和类别的简笔画经过编码，被集中投影到 GMM 中的一个高斯区域中。通过选取不同高斯中的编码，可以可控地生成具有指定特征的简笔画。然而，现有方法在处理形态相似的简笔画时，所构建的 GMM 空间中，高斯区域间存在较大重叠。这降低了简笔画生成符合预期特征的准确率，即可控生成性能较差。本文以贝叶斯阴阳和谐学习算法为指导提出了 IA-pix2seq 深度双向学习模型。模型的双向互逆映射在和谐学习原理指导下，以最默契的方式达到最大共识，将同一高斯成分区域内的编码集中到相应的高斯中心，同时进一步约束了各简笔画在编码空间中的投影范围，从而扩大高斯成分间的边界并降低彼此间的重叠率。实验表明 IA-pix2seq 能有效降低不同类别简笔画因相似造成的编码重叠，以提高简笔画的可控生成性能。给定插值编码、将含像素缺失的简笔画作为约束，模型生成的简笔画仍能保留更多的预期特征。

关键词　简笔画生成; 编码自组织; 贝叶斯阴阳和谐学习; 深度双向智能系统; 高斯混合模型

中图法分类号　TP391

IA-pix2seq: A Deep Bidirectional Learning Method for Controllable Sketch Synthesis

ZANG Si-Cong　TU Shi-Kui　XU Lei

(Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240)

Abstract　Drawing free-hand sketches is a simple way for communication and expression in human history. A free-hand sketch carries vivid messages and emotions from its flexible drawing manner. Thus, sketches are always abstract, lack-of-details and variant. Synthesizing a sketch to follow the expected plan requires controlling both the categorial sketch pattern and the non-categorical one, i.e., the stylistic pattern. But the labels of stylistic pattern are not given in the sketch datasets, leaving the controllable sketch synthesis to be an unsupervised task. A group of recent studies target to build a latent space, preserving the similarity of structural patterns from the observed sketch data to the latent codes. Such a latent space is regarded as a Voronoi tessellation, where each latent region is assigned a unique concept or a pattern, representing a specific sketch category and style. Thus, it is practical to locate a specific latent code to synthesize a sketch with the expected patterns. A simple approach is self-organizing a Gaussian mixture model (GMM) distributed latent space, where each Gaussian component in GMM represents a specific sketch pattern. However, these Gaussians are heavily overlapped in recent studies when facing sketches with similar structural patterns. The controllable synthesis performance drops due to the tiny margins between latent regions. We present yIng-yAng system pixel to

本课题得到国家科技部科技创新2030-新一代人工智能重大项目(No.2018AAA0100700)、上海市科委人工智能重大项目(No.2021SHZDZX0102)的资助.
臧思聪, 博士研究生, 主要研究领域为机器学习. E-mail: SCZang@sjtu.edu.cn. **涂仕奎（通信作者）**, 博士，副教授，中国计算机学会(CCF)会员(80996M)，主要研究领域为机器学习、生物信息学. E-mail: tushikui@sjtu.edu.cn. **徐 雷（通信作者）**，博士，教授，主要研究领域为机器学习、因果计算、双向智能、类AlphaGo系统、智能医疗、智慧金融. E-mail: leixu@sjtu.edu.cn.

sequence (IA-pix2seq) guided by Bayesian Ying-Yang (BYY) harmony learning algorithm. The structure of IA-pix2seq can be divided into two levels. The bottom level contains a convolutional neural network (CNN) encoder, a recurrent neural network (RNN) decoder and a CNN decoder, which are designed for feature extraction, sketch generation and regularization, respectively. The deep network optimization in bottom level is guided by maximizing the harmony measure from BYY learning. The top level is for updating the parameters of latent GMM distribution, and BYY learning is also utilized to solve this GMM learning task. During training, a mini-batch of sketches are fed into the bottom level to obtain their corresponding latent codes. These codes are sent into the top level to update the GMM latent distribution, which further regularizes the bottom network to produce sketch codes fitting to the current GMM distribution. Finally, the latent codes are sent into the decoder for reconstructing the same sketches as input, ensuring the detailed sketch features are correctly embedded into the latent codes. BYY harmony learning seeks a best matching between encoding and decoding subsystems with a most tacit manner by minimizing the information transferred from sketches to latent codes. Correspondingly, IA-pix2seq not only centralizes the latent codes within a latent Gaussian component but also squeezes the latent territory for each sketch sample. As a result, wide margins are left between latent regions, contributing to the reduced overlaps and further a more easily controlled sketch synthesis process. Experimental results show that IA-pix2seq improves the controllable synthesis performance especially on sketches with similar structural patterns or with a variety of styles. The generated sketches preserve more expected details from the input constraint which can be either an interpolated latent code or a masked sketch image.

Key words sketch synthesis; latent codes self-organization; Bayesian Ying-Yang harmony learning; deep bidirectional intelligent system; Gaussian mixture model

1 引 言

简笔画作为一种信息传递和感情表达的媒介，在人类认知发展进程中具有重要意义。其往往呈现抽象、多变、缺乏细节而又不乏生动的特点。即使是表示同一物体类别的简笔画，也会由于绘画者绘制方式（如是否一笔绘制完成简笔画）、概念偏差（如图 1 中的简笔画“猪”，既可以仅用脑袋，也可以用带完整身体的形象表达）的不同，造成简笔画视觉形态的巨大差异。文献[1]将这些非类别层面的多样性统称为“风格”。这里的“风格”类似手写数字中笔画的粗细、倾斜程度[2]，或是图像艺术风格迁移（Style Transfer）[3]中的艺术形态等。而简笔画的可控生成，需同时考虑其类别和风格两方面。

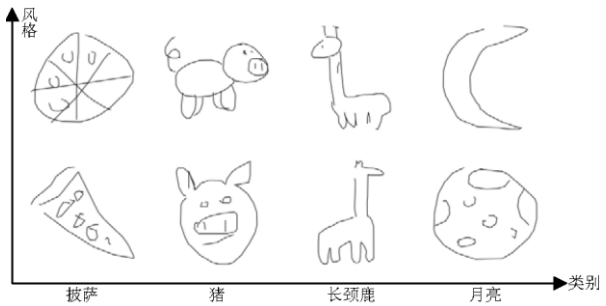

图 1 同一类别的简笔画存在截然不同的风格

据我们所知，现有的简笔画数据集，如 QuickDraw[4]、Sketchy database[5]和 TU-Berlin[6]，并不包含与风格特征相关的标签。风格标签的缺失使得简笔画生成模型无法通过有监督学习去训练。文献 [1] 假设简笔画的编码服从高斯混合模型（Gaussian Mixture Model, GMM）分布，并无监督地将每一种特定类别和风格的简笔画投影到 GMM 的一个高斯成分中。由此，通过选取不同高斯中的编码，实现可控地生成具有指定特征的简笔画。然而 RPCL-pix2seq 构建的 GMM 编码空间中，各高斯成分存在较大重叠。尤其在处理形态相似的简笔画样本时，重叠更严重，可能造成高斯数量 K 的误判。如果选取重叠区域的编码去生成简笔画，可能导致随机产生不同类别与风格的简笔画。若能更紧凑地组织 GMM 各高斯成分内的编码，拓宽高斯间的边界，可以提升简笔画的可控生成性能。

为此，本文提出了 yIng-yAng system pixel to sequence（IA-pix2seq）模型。该模型从外向内将简笔画映射为编码，从内向外解码重建或生成简笔画，是一种深度双向智能系统[7][8]。使用贝叶斯阴阳和谐学习（Bayesian Ying-Yang (BYY) Harmony Learning）[9][10]作为模型训练准则，BYY 和谐学习将 IA-pix2seq 的双向互逆映射以最默契的方式达到最大共识，不仅寻求从内部编码到外部很好地重建或生成简笔画，而且同时从外向内的编码过程是其最优的逆，并且促使整个双向系统复杂度尽量小。

具体地，IA-pix2seq 不仅能将相同类别和风格的简笔画编码形成紧凑的聚簇，而且能鲁棒地区分形态相似、类别不同的简笔画，保证同一聚簇中简笔画特征的唯一性和一致性，降低编码重叠率，提高简笔画的可控生成性能。

本文的主要贡献总结如下：1）提出 IA-pix2seq 模型、并通过 BYY 和谐学习准则得到相应的学习算法。该算法可以降低不同特征简笔画的编码重叠率，有效提高可控生成性能；2）实验结果显示了 IA-pix2seq 模型在自组织编码、可控生成方面的优越性。而且，平滑紧凑的编码空间使得 IA-pix2seq 可以高效地生成新颖的、有创意的简笔画，还能对像素缺失的简笔画图片补全细节。

本文算法的源代码以及更多的实验结果将会公布在：https://github.com/CMACH508/IA-pix2seq。

2 相关工作

2.1 简笔画生成模型的编码空间结构

作为最早提出的简笔画生成方法之一，sketch-rnn[4]假设简笔画编码服从标准正态分布。不同类别、风格的简笔画编码被紧凑地约束在同一个高斯分布中。我们很难从这样一个紧密而杂乱的编码空间中选取一个合适的编码，使其生成具有期望特征的简笔画。因此，文献[11]指出，sketch-rnn 在同时处理多类别简笔画时表现较差。

为此，sketch-pix2seq[11]在目标函数中去掉了 Kullback-Leibler（KL）散度项以放松对编码的单高斯约束。更松弛的编码空间结构鼓励相似类别、风格的简笔画投影到各自的编码区域，从而提升在多类别数据集上的可控生成性能。然而过于松弛的编码分布使得表示特定类别、风格的编码区域内部不够紧凑，区域间的边界模糊，重叠严重。如果选取的编码来自重叠区，解码生成的可能是随机呈现的不同类别、不同风格的简笔画，可控性差。

构造一个 GMM 分布的编码空间，并为 GMM 的每个高斯成分指派一种特定类别、风格的简笔画编码，可以加强相同特征简笔画编码间的聚合，形成相应的高斯聚簇。RPCL-pix2seq[1]构造了一个分层结构的网络，将底层网络参数和顶层 GMM 参数的训练过程分离：1）网络顶层通过结合对手惩罚竞争学习（Rival Penalized Competitive Learning, RPCL）[12]的期望最大化（Expectation-Maximization, EM）[13][14]算法估计 GMM 参数，2）底层通过网络优化器更新网络参数。通过引入 RPCL 策略，RPCL-pix2seq 根据简笔画数据特征自动确定 GMM 的高斯成分数量。这使得编码空间的 GMM 结构更为准确、鲁棒，避免因 GMM 高斯成分数不准确，引入冗余或杂乱的编码聚簇。

然而，在面对形态相近的不同类别简笔画时，RPCL-pix2seq 构建的对应不同特征的高斯成分编码区域之间仍然存在较大重叠，影响可控生成的性能。本文借鉴了这种分层的网络结构，引入 BYY 和谐学习[9][10]指导网络顶层的 GMM 参数学习，进一步解决编码区重叠问题。

2.2 BYY和谐学习简介

BYY 和谐学习[9][10]是一个统一的统计学习理论。整个学习系统包含外部观测域 $\boldsymbol{X}$（如数据空间）和其对应的内部表达域 $\boldsymbol{R}$（如编码空间）两部分。$\boldsymbol{X}$ 和 $\boldsymbol{R}$ 的联合概率可以分解为以下两种等价形式：Ying 模型 $q(\boldsymbol{X}|\boldsymbol{R})q(\boldsymbol{R})$，含由 $\boldsymbol{R}$ 至 $\boldsymbol{X}$ 方向的 Ying 通道 $q(\boldsymbol{X}|\boldsymbol{R})$ 和编码分布 $q(\boldsymbol{R})$；Yang 模型 $p(\boldsymbol{R}|\boldsymbol{X})p(\boldsymbol{X})$ 作为 Ying 模型的逆，含由 $\boldsymbol{X}$ 至 $\boldsymbol{R}$ 方向的 Yang 通道 $p(\boldsymbol{R}|\boldsymbol{X})$ 和数据分布 $p(\boldsymbol{X})$。BYY 系统的学习目标是最大化和谐函数 $H(p\|q)$，即

$$\begin{aligned}&\max H(p\|q)\\&=\int p(\boldsymbol{R}|\boldsymbol{X})p(\boldsymbol{X})\log q(\boldsymbol{X}|\boldsymbol{R})q(\boldsymbol{R})d\boldsymbol{X}d\boldsymbol{R}.\end{aligned}\tag{1}$$

公式(1)可等价为

$$H(p\|q)=H(p\|p)-KL(p\|q).\tag{2}$$

公式(2)中的 $H(p\|p)$ 表示Yang模型 $p(\boldsymbol{R}|\boldsymbol{X})p(\boldsymbol{X})$ 的负熵，最大化 $H(p\|p)$ 表示最小化Yang模型从 $\boldsymbol{X}$ 到 $\boldsymbol{R}$ 方向上的编码复杂度；$KL(p\|q)$ 表示Ying和Yang两个子系统间的KL散度，最大化 $-KL(p\|q)$ 迫使 $p(\boldsymbol{X},\boldsymbol{R})$ 与 $q(\boldsymbol{X},\boldsymbol{R})$ 形成最佳匹配。整体而言，和谐原理使得Ying和Yang以交换信息最小的方式达到最大共识，不仅追求Ying能很好地描述数据 $\boldsymbol{X}$，而同时Yang作为Ying的最优逆，并促使BYY系统复杂度尽量小[8][9]。

深度双向智能系统[8]是上述 BYY 系统在深度学习上的一个应用。Ying 模型和 Yang 模型可通过深度神经网络实现，得到从内向外的 I-映射、从外向内的 A-映射。

3 方法

3.1 IA-pix2seq概述

IA-pix2seq 假设简笔画编码服从 GMM 分布，不同高斯成分表示简笔画的各自类别和风格。具体地，编码 $\boldsymbol{y}$ 的分布为

$$\begin{aligned}q(\boldsymbol{y})&=\mathrm{GMM}(\boldsymbol{y}|K,\boldsymbol{\Theta}_{k=1}^{K})\\&=\sum_{k=1}^{K}\alpha_k G(\boldsymbol{y}|\boldsymbol{\mu}_k,\mathrm{diag}(\boldsymbol{\sigma}_k^2)).\end{aligned}\tag{3}$$

式中超参数 K 表示 GMM 中高斯成分的数量；$\boldsymbol{\Theta}_k=\{\alpha_k,\boldsymbol{\mu}_k,\mathrm{diag}(\boldsymbol{\sigma}_k^2)\}$ 对应第 k 个高斯成分的混合概率（$\alpha_k\geq 0$，$\sum_{k=1}^{K}\alpha_k=1$）、均值和对角化协方差矩阵。为简化计算，本文假设 GMM 参数 $q(\boldsymbol{\Theta}_k)$ 的先验分布是一个 Dirac δ 函数。

一个简笔画 $\boldsymbol{X}$ 的内部表达 $\boldsymbol{R}$ 既包含一个矢量

编码 $\boldsymbol{y}$，又包括 $\boldsymbol{y}$ 对应的高斯成分编号 k。本文构造了一种分层的映射方式 $\boldsymbol{X}\to \boldsymbol{y}\to k$，如下：

$$p(\boldsymbol{X},\boldsymbol{y},k)\Rightarrow \underbrace{p(\boldsymbol{x}_i\mid \boldsymbol{X})}_{①\text{ Mini-batching}}\cdot \underbrace{p(\boldsymbol{y}\mid \boldsymbol{x}_i)}_{②\text{ Encoding}}\cdot \underbrace{p(k\mid \boldsymbol{y})}_{③\text{ Assigning}}, \quad (4)$$

$$q(\boldsymbol{X},\boldsymbol{y},k)\Rightarrow \underbrace{q(k)}_{④\text{ Selecting}}\cdot \underbrace{q(\boldsymbol{y}\mid k)}_{⑤\text{ Sampling}}\cdot \underbrace{q(\boldsymbol{x}_i\mid \boldsymbol{y})}_{⑥\text{ Generating}}. \quad (5)$$

公式(4)中，$p(\boldsymbol{X},\boldsymbol{y},k)$ 分解为：①从训练集 $\boldsymbol{X}$ 中随机抽取小批次（Mini-batch）的简笔画。②将抽取的简笔画 $\boldsymbol{x}_i$ 送入编码器 $p(\boldsymbol{y}\mid \boldsymbol{x}_i)$ 获得相应的编码 $\boldsymbol{y}$。③计算编码 $\boldsymbol{y}$ 对第 k 个高斯成分的指派 $p(k\mid \boldsymbol{y})$。与文献[1][15]相同，本文假设编码 $\boldsymbol{y}$ 的指派过程不直接依赖于 $\boldsymbol{X}$，即 $p(k\mid \boldsymbol{X},\boldsymbol{y})=p(k\mid \boldsymbol{y})$。公式(5)中，$q(\boldsymbol{X},\boldsymbol{y},k)$ 则分解为：④从 GMM 中选择一个高斯成分 $G(\boldsymbol{y}\mid \boldsymbol{\mu}_k,\mathrm{diag}(\boldsymbol{\sigma}_k^2))$。⑤从选取的高斯 G_k 中抽取编码 $\boldsymbol{y}$。⑥将 $\boldsymbol{y}$ 送入解码器生成相应的简笔画 $\boldsymbol{x}_i$。本文同样假设简笔画的生成过程不直接取决于指派 k，即 $q(\boldsymbol{x}_i\mid \boldsymbol{y},k)=q(\boldsymbol{x}_i\mid \boldsymbol{y})$。

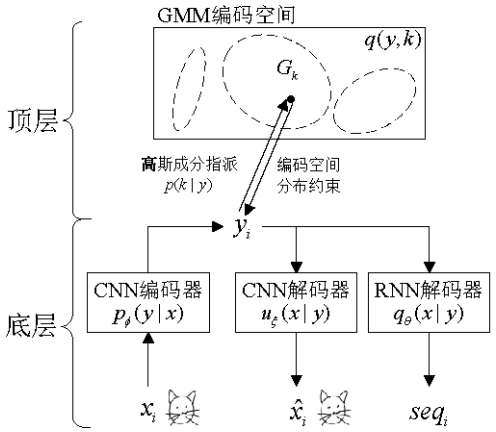

图 2 IA-pix2seq 的概览图

由此，本文构建了如图 2 的双向网络结构。IA-pix2seq 的底层是一个深度编、解码器，自下而上从简笔画图片 $\boldsymbol{x}_i$ 提取特征并编码，自上而下解码重建图片格式的简笔画 $\hat{\boldsymbol{x}}_i$、生成序列格式的简笔画 seq_i。底层将顶层的 GMM 分布作为编码的先验约束代入目标函数，通过网络优化器的梯度反传完成底层参数 $\boldsymbol{\phi}$、$\boldsymbol{\xi}$ 和 $\boldsymbol{\theta}$ 更新。顶层通过 BYY 和谐学习算法，根据底层提取的编码 $\boldsymbol{y}$ 更新 GMM 参数 $\boldsymbol{\Theta}_{k=1}^{K}$。

需要注意的是，IA-pix2seq 的顶层输入 $\boldsymbol{y}$ 是动态的。首先，每次训练输入的 $\boldsymbol{x}_i$ 是小批次的，即不同批次参与训练的 $\boldsymbol{y}$ 不同；其次，训练时编码器参数 $\boldsymbol{\phi}$ 的更新导致编码映射 $p_\phi(\boldsymbol{y}\mid \boldsymbol{x}_i)$ 不断变化，相同简笔画 $\boldsymbol{x}_i$ 在不同训练时刻所对应的编码 $\boldsymbol{y}$ 也不同。

3.2 IA-pix2seq底层：简笔画的特征抽取和生成

IA-pix2seq 底层由一个卷积神经网络（Convolutional Neural Network, CNN）编码器、一个循环神经网络（Recurrent Neural Network, RNN）解码器和一个 CNN 解码器组成。输入为 48×48 像素大小的简笔画图片，由 QuickDraw 数据集的原始序列格式转换得到[1][11]。CNN 编码器保持与 RPCL-pix2seq 中相一致的结构。RNN 解码器采用 sketch-rnn 中的 hyper LSTM[16]。与 RPCL-pix2seq 相同，IA-pix2seq 也引入了一个 CNN 解码器，通过重建输入对网络参数更新提供正则化约束，最大程度保证学习得到的编码空间结构同真实简笔画分布相一致。

将公式(4)-(5)代入公式(1)中的和谐函数，得到 IA-pix2seq 底层训练的目标函数

$$\begin{aligned}&\max_{\phi,\xi,\theta} L(\phi,\xi,\theta;\boldsymbol{x}_i,\boldsymbol{\Theta})\\&=\lambda_1\mathrm{E}_{p_\phi(y\mid x_i)}[\log q_\theta(\boldsymbol{x}_i\mid \boldsymbol{y})]\\&\quad+\lambda_2\mathrm{E}_{p_\phi(y\mid x_i)}[\log u_\xi(\boldsymbol{x}_i\mid \boldsymbol{y})]\\&\quad+\mathrm{E}_{p_\Theta(k\mid y)p_\phi(y\mid x_i)}[\log q_\Theta(\boldsymbol{y}\mid k)q_\Theta(k)].\end{aligned} \quad (6)$$

式中的 λ_1 和 λ_2 为超参数加权（$\lambda_1,\lambda_2>0$）。其中第一项 $\mathrm{E}_{p_\phi(y\mid x_i)}[\log q_\theta(\boldsymbol{x}_i\mid \boldsymbol{y})]$ 来自生成序列格式简笔画的分支，采用 sketch-rnn 中的计算方法实现最大化；第二项 $\mathrm{E}_{p_\phi(y\mid x_i)}[\log u_\xi(\boldsymbol{x}_i\mid \boldsymbol{y})]$ 来自图片格式的分支，使用最小均方误差（Mean Square Error, MSE）$\|\boldsymbol{x}_i-\hat{\boldsymbol{x}}_i\|^2$ 计算；第三项来自于对编码的 GMM 分布约束。同时，本文假设 $p_\phi(\boldsymbol{y}\mid \boldsymbol{x})=G(\boldsymbol{y}\mid \boldsymbol{a},\mathrm{diag}(\boldsymbol{b}^2))$，$\boldsymbol{a}$ 和 $\mathrm{diag}(\boldsymbol{b}^2)$ 分别表示高斯分布的均值和对角化协方差矩阵。通过重参数化[17] $\boldsymbol{y}=\boldsymbol{a}+\boldsymbol{b}\odot\boldsymbol{\varepsilon}$ 得到编码 $\boldsymbol{y}$，其中 $\boldsymbol{a}$、$\boldsymbol{b}$ 由 CNN 编码器直接计算得到，$\odot$ 表示 Hadamard 积，$\boldsymbol{\varepsilon}$ 从标准正态分布 $G(\boldsymbol{\varepsilon}\mid \boldsymbol{0},\boldsymbol{I})$ 中随机抽样获得，其中 $\boldsymbol{0}$ 和 $\boldsymbol{I}$ 分别表示零向量和单位矩阵。由此，我们可计算

$$\begin{aligned}&\mathrm{E}_{p_\Theta(k\mid y)p_\phi(y\mid x_i)}[\log q_\Theta(\boldsymbol{y}\mid k)q_\Theta(k)]\\&=-\sum_k\frac{p_\Theta(k\mid \boldsymbol{y})}{2}\sum_{j=1}^{D}\left[\log 2\pi+\log\sigma_j^2+\frac{b_{ij}^2+(a_{ij}-\mu_j)^2}{\sigma_j^2}\right]\\&\quad-\sum_k p_\Theta(k\mid \boldsymbol{y})\log\alpha_k.\end{aligned} \quad (7)$$

式中 D 表示编码 $\boldsymbol{y}$ 的维度；$p_\Theta(k\mid \boldsymbol{y})$ 表示编码 $\boldsymbol{y}$ 对应 GMM 各高斯成分的指派，在 IA-pix2seq 的顶层计算。

3.3 IA-pix2seq顶层：GMM编码空间的自组织

IA-pix2seq 顶层的 GMM 参数 $\boldsymbol{\Theta}$ 学习通过 BYY 和谐学习的 Ying-Yang 迭代实现[18][19]。

Yang 步骤：

$$p^{(t)}(k\mid \boldsymbol{y}_i)=\frac{\alpha_k^{(t-1)}\cdot G\left(\boldsymbol{y}_i\mid \boldsymbol{\mu}_k^{(t-1)},\mathrm{diag}(\boldsymbol{\sigma}_k^{2(t-1)})\right)}{\sum_j \alpha_j^{(t-1)}\cdot G\left(\boldsymbol{y}_i\mid \boldsymbol{\mu}_j^{(t-1)},\mathrm{diag}(\boldsymbol{\sigma}_j^{2(t-1)})\right)},$$
$$\delta_{ik}^{(t)}=\log p^{(t)}(k\mid \boldsymbol{y}_i)-\sum_j p^{(t)}(j\mid \boldsymbol{y}_i)\log p^{(t)}(j\mid \boldsymbol{y}_i),$$
$$\tilde{p}_{ik}^{(t)}=p^{(t)}(k\mid \boldsymbol{y}_i)\cdot(1+\delta_{ik}^{(t)}). \tag{8}$$

Ying 步骤：

$$\tilde{\alpha}_k^{(t)}=\tilde{q}^{(t)}(k)=\sum_{i=1}^N \tilde{p}_{ik}^{(t)}\Big/N,\ \tilde{\boldsymbol{\mu}}_k^{(t)}=\frac{\sum_i \tilde{p}_{ik}^{(t)}\cdot \boldsymbol{a}_i}{\sum_i \tilde{p}_{ik}^{(t)}},$$
$$\mathrm{diag}(\tilde{\boldsymbol{\sigma}}_k^{2(t)})=\frac{\sum_i \tilde{p}_{ik}^{(t)}\cdot[(\boldsymbol{a}_i-\tilde{\boldsymbol{\mu}}_k^{(t)})(\boldsymbol{a}_i-\tilde{\boldsymbol{\mu}}_k^{(t)})^{\mathrm{T}}+\mathrm{diag}(\boldsymbol{b}_i^2)]}{\sum_i \tilde{p}_{ik}^{(t)}}. \tag{9}$$

参数更新：

$$\begin{aligned}
&\boldsymbol{\mu}_k^{(t)}=(1-\eta)\boldsymbol{\mu}_k^{(t-1)}+\eta\tilde{\boldsymbol{\mu}}_k^{(t)},\\
&\boldsymbol{\sigma}_k^{2(t)}=\left[\sigma_{k1}^{2(t)},\sigma_{k2}^{2(t)},\cdots,\sigma_{kD}^{2(t)}\right],\\
&\sigma_{kd}^{2(t)}=\max\{(1-\eta)\sigma_{kd}^{2(t-1)}+\eta\tilde{\sigma}_{kd}^{2(t)},10^{-5}\},\\
&\alpha_k^{(t)}=\max\{(1-\eta)\alpha_k^{(t-1)}+\eta\tilde{\alpha}_k^{(t)},0\}.
\end{aligned} \tag{10}$$

在第(t)时刻的训练中，底层特征提取得到的编码 $\{\boldsymbol{y}_i\}_{i=1}^N$ 被送入顶层，结合第(t-1)时刻的 GMM 参数 $\boldsymbol{\Theta}_k^{(t-1)}=\{\alpha_k^{(t-1)},\boldsymbol{\mu}_k^{(t-1)},\mathrm{diag}(\boldsymbol{\sigma}_k^{2(t-1)})\}$，根据公式(8)中的 Yang 步骤，计算编码 $\boldsymbol{y}_i$ 在当前第 k 个高斯成分的指派 $\tilde{p}_{ik}^{(t)}$。接着，$\tilde{p}_{ik}^{(t)}$ 被送入公式(9)中的 Ying 步骤，结合公式(10)中的学习率 η 完成当前第(t)时刻 GMM 参数 $\boldsymbol{\Theta}^{(t)}$ 的更新。其中，公式(10)中的 max 操作逐个元素地取相应元素与 10^{-5} 或 0 的最大值，为确保协方差矩阵 $\mathrm{diag}(\boldsymbol{\sigma}_k^{2(t)})$ 的正定性、以及混合概率 $\alpha_k^{(t)}$ 的非负性。

公式(8)中的 δ_{ik} 计算的是第 k 个高斯成分 G_k 对编码 $\boldsymbol{y}_i$ 的相对拟合度，是 BYY 和谐学习与最大似然 EM 算法的区别所在。如果固定 $\delta_{ik}=0$，上述更新步骤退回到 EM 算法。此外，

- $\delta_{ik}>0$：参数更新方向与最大似然的方向一致，而且增大了更新的力度；
- $-1\le\delta_{ik}<0$：参数更新方向仍然与最大似然的方向一致，但是削弱了更新的力度；
- $\delta_{ik}<-1$：参数更新方向与最大似然方向相反，此时类似 RPCL 算法中为对手施加的逆学习。BYY 和谐学习可以同时向多个对手施加逆学习，且逆学习强度由和谐函数自动确定，表现应优于 RPCL。RPCL 是 BYY 和谐学习的一个特殊情况。

基于上述算法细节，IA-pix2seq 在训练过程中不仅可以自动确定 GMM 编码分布的高斯成分数 K，而且得到的编码分布较之于 RPCL-pix2seq 更紧凑。特别地，IA-pix2seq 更能在编码空间区分形态相似但类别和风格不同的简笔画。

从算法原理角度分析，根据 IA-pix2seq 对公式(4)中 $p(\boldsymbol{X},\boldsymbol{y},k)$ 的分解，和谐学习较最大似然学习（Maximum Likelihood Learning, ML）的区别体现在最大化公式(2)中的 $H(p\|p)$ 项，即

$$\begin{aligned}
&\max \int p_\phi(\boldsymbol{y}\mid \boldsymbol{x}_i)\log p_\phi(\boldsymbol{y}\mid \boldsymbol{x}_i)d\boldsymbol{y}\\
&+\int p_\phi(\boldsymbol{y}\mid \boldsymbol{x}_i)\sum\nolimits_k p_{\boldsymbol{\Theta}}(k\mid \boldsymbol{y})\log p_{\boldsymbol{\Theta}}(k\mid \boldsymbol{y})d\boldsymbol{y}.
\end{aligned} \tag{11}$$

最大化公式(11)中的第一项等价于减小 $p_\phi(\boldsymbol{y}\mid \boldsymbol{x}_i)=G(\boldsymbol{y}\mid \boldsymbol{a}_i,\mathrm{diag}(\boldsymbol{b}_i^2))$ 中的 $\sum_{d=1}^D\log \boldsymbol{b}_i^2$，即压缩各简笔画样本在编码空间中占据的投影区域。而对于公式(11)中的第二项，鼓励 $\{p_{\boldsymbol{\Theta}}(k\mid \boldsymbol{y})\}_{k=1}^K$ 逼近独热编码（One-hot）形式，即进一步增大 $p_{\boldsymbol{\Theta}}(k\mid \boldsymbol{y})$ 的最大指派。一方面，通过调整编码器参数 ϕ 实现相同类别、相似风格简笔画在编码空间中的集聚；另一方面，更新 GMM 参数 $\boldsymbol{\Theta}$ 进一步扩大编码 $\boldsymbol{y}$ 在 G_k 上的指派值，以提高各高斯成分内部的紧凑性。

综上，IA-pix2seq 在构建 GMM 分布的编码空间时，保证了各高斯内部的紧凑性，降低了高斯间的重叠率，还可以自动剔除 GMM 中的冗余高斯。

4 实验与分析

4.1 实验准备

为验证模型性能，本文利用 QuickDraw[4]的简笔画构造了三个数据集。数据集 1 选用文献[1]使用的简笔画类别，以方便与该文献的实验结果直接对比。该数据集所含同类别的简笔画存在差异明显的风格。数据集 2 选用两个形态相似的简笔画类别，以直接比较在简笔画细节上的认知把控能力。数据集 3 集合了更多类别的简笔画，且不同类别间存在一致的风格，进一步增大可控生成难度。具体设置如下：

数据集 1：含蜜蜂、公交车、花、长颈鹿和猪共 5 个类别，合计 350K（1K=10^3）个训练样本、12.5K 个验证样本和 12.5K 个测试样本。

数据集 2：含猫和猪共 2 个类别，合计 140K 个训练样本、5K 个验证样本和 5K 个测试样本。猫和猪的最大区别仅在于猫的胡须和猪的鼻子。

数据集 3：在数据集 1 上增加汽车、猫和马共 3 个类别，合计 560K 个训练样本、20K 个验证样本和 20K 个测试样本。长颈鹿和马都拥有向左/右朝向的风格特征；猫和猪都拥有一致的面部结构；汽车和公交车都大致呈现左右对称的形态。

本节实验比较本文提出的 IA-pix2seq 与 sketch-rnn[4]、sketch-pix2seq[11]、Song et al.[20]、VaDE[15]、SketchHealer[21]和 RPCL-pix2seq[1]的简笔画可控生成性能。我们为 Song et al.和 VaDE 提供了

与 IA-pix2seq 相同结构的编码器和解码器，以适配 QuickDraw 数据集。同时，遵照文献[15]为 VaDE 提供了 1 个 epoch 的预训练。对于使用 GMM 编码空间的 VaDE、RPCL-pix2seq 和 IA-pix2seq，在三个数据集下的高斯数量 K 分别初始化为 10、8 和 16。另，我们为 IA-pix2seq 替换了 SketchHealer 的编码器以借鉴其特征提取方法，相应模型命名为 IA-pix2seq+，同样参与实验比较。

为量化简笔画生成可控性能，本文在整个测试集上计算 *Rec* [1]、*Ret* [1]和 *Acc* [15]。*Rec* 描述生成简笔画的类别是否可控，我们针对三个数据集分别训练 sketch-a-net[22]作为分类器以计算 *Rec*； *Ret* 描述生成简笔画的风格是否可控；对于 GMM 分布的编码空间，*Acc* 描述不同简笔画类别是否准确指派至不同的高斯成分。具体计算如下：

- *Rec*：用简笔画 $\boldsymbol{x}$ 的编码通过解码器生成简笔画 $\hat{\boldsymbol{x}}$。*Rec* 表示 $\hat{\boldsymbol{x}}$ 与 $\boldsymbol{x}$ 属于同一类别的概率。*Rec* 越大，表示模型的特征抽取能力、解码生成性能越好；
- *Ret*：记 $\boldsymbol{y}$ 是简笔画 $\boldsymbol{x}$ 的编码，以 $\boldsymbol{y}$ 为输入解码生成得到简笔画 $\hat{\boldsymbol{x}}$，将 $\hat{\boldsymbol{x}}$ 输入编码器得到 $\hat{\boldsymbol{y}}$。若 $\hat{\boldsymbol{y}}$ 与 $\boldsymbol{y}$ 之间的距离很小，那么说明模型生成的 $\hat{\boldsymbol{x}}$ 很好保持了原简笔画 $\boldsymbol{x}$ 的类别和风格特征。我们将测试集中所有的简笔画送入编码器得到的编码集 $\boldsymbol{Y}$ 作为对照，从 $\boldsymbol{Y}$ 中检索与 $\hat{\boldsymbol{y}}$ 距离最近的编码。*Ret* 表示成功检索出 $\boldsymbol{y}$ 的概率。*Ret* 越大，表示模型生成的简笔画与预期保持类别和风格一致的能力越强；
- *Acc*：简笔画被投影到与之类别相一致的 GMM 高斯成分中的准确率，计算公式如下，

$$Acc=\max_{m\in M}\frac{\sum_{i=1}^{N}1\{l_i=m(c_i)\}}{N}.$$

式中的 M 表示高斯成分和类别标签间的一一映射，l_i 和 c_i 分别表示测试集中第 i 个简笔画的真实类别标签和其在高斯成分上的指派。*Acc* 越大，表示模型构造的编码空间中不同类别的区域划分越准确。

实验使用一块 NVIDIA Tesla P40 GPU 训练模型。批大小（Mini-batch Size）N、编码维度 D、GMM 参数 $\boldsymbol{\Theta}$ 的学习率 η 和目标函数中权值 λ_1，分别设置为 400、128、0.05 和 1.0。权值 λ_2 初始化为 0.5，但从第 4 个训练 epoch 起调整为 0.01。网络训练使用 Adam 优化器，初始学习率设置为 0.001，而后每训练 1 个 epoch 衰减为当前的 95%。Adam 优化器的超参数分别设置为 $\beta_1=0.05$、$\beta_2=0.999$。

4.2 简笔画生成结果的可控性比较

表 1-表 3 分别给出了各方法在三个数据集下的可控生成指标。过于紧凑的单高斯编码分布，限制了 sketch-rnn 在多类简笔画数据下的可控生成性能。而在面对类别数较少的数据集 2 时，sketch-rnn 的可控性能优于其在数据集 1 和数据集 3 下的表现。sketch-pix2seq 和 Song et al.分别通过松弛编码空间约束，或是引入多种简笔画表达形式的方式，鼓励不同类别和风格的简笔画编码形成各自鲜明的编码区域，在三个数据集下较 sketch-rnn 均有不同程度的性能提升。SketchHealer 吸收了两方法的优点，不仅去除目标函数中的 KL 散度项，还使用图结构丰富了简笔画图片所缺失的简笔画绘制顺序，提高了模型的可控生成性能。VaDE 和 RPCL-pix2seq 将相同特征的简笔画编码约束在相应的 GMM 高斯成分中。同时，RPCL-pix2seq 的对手惩罚机制可以自动确定一个合适的高斯成分数，尽可能维持各高斯与简笔画类别、风格间的一一对应。更准确的 GMM 编码结构保证 RPCL-pix2seq 具有相对优异和稳定的可控生成性能。

表 1 数据集 1 下简笔画可控生成性能比较（%）

方法	*Rec* ↑	*Ret* ↑			*Acc* ↑
		Top-1	Top-10	Top-50	
sketch-rnn[4]	50.33	0.38	2.84	9.33	/
sketch-pix2seq[11]	83.99	13.45	30.12	49.99	/
Song et al.[20]	91.77	16.41	36.43	52.22	/
VaDE[15]	87.83	3.30	12.59	26.34	57.87
RPCL-pix2seq[1]	93.18	17.86	38.87	55.30	73.60
IA-pix2seq	**94.75**	24.34	48.08	67.36	93.31
SketchHealer[21]	91.04	58.80	82.15	91.33	/
IA-pix2seq+	93.02	**64.09**	**86.59**	**94.25**	**96.41**

表 2 数据集 2 下简笔画可控生成性能比较（%）

方法	*Rec* ↑	*Ret* ↑			*Acc* ↑
		Top-1	Top-10	Top-50	
sketch-rnn[4]	80.66	17.40	42.60	54.52	/
sketch-pix2seq[11]	79.74	16.64	39.50	57.20	/
Song et al.[20]	81.54	20.56	42.08	58.26	/
VaDE[15]	74.68	6.14	22.92	41.72	59.00
RPCL-pix2seq[1]	82.08	18.56	42.98	58.02	66.40
IA-pix2seq	82.74	24.38	50.18	68.54	81.12
SketchHealer[21]	85.84	49.36	74.26	87.36	/
IA-pix2seq+	**87.00**	**59.04**	**80.56**	**91.10**	**83.03**

表 3 数据集 3 下简笔画可控生成性能比较（%）

方法	*Rec* ↑	*Ret* ↑			*Acc* ↑
		Top-1	Top-10	Top-50	
sketch-rnn[4]	57.64	3.72	13.42	26.14	/
sketch-pix2seq[11]	79.13	22.92	47.55	58.19	/
Song et al.[20]	83.28	25.47	43.39	56.16	/
VaDE[15]	74.68	4.78	19.22	39.51	43.68
RPCL-pix2seq[1]	81.80	28.80	59.05	77.52	50.62
IA-pix2seq	83.99	37.28	58.09	83.53	76.50
SketchHealer[21]	87.03	68.52	82.37	86.57	/
IA-pix2seq+	**88.09**	**80.01**	**93.69**	**97.51**	**85.42**

然而，RPCL-pix2seq 所构造的 GMM 结构中，多数高斯成分重叠严重。尤其在面对简笔画形态更为相似的数据集 2 和数据集 3 时，从重叠区域中抽取的编码，很难准确对应与期望一致的简笔画特征。图 3 给出了 RPCL-pix2seq 和 IA-pix2seq 所构造编码空间的可视化比较。

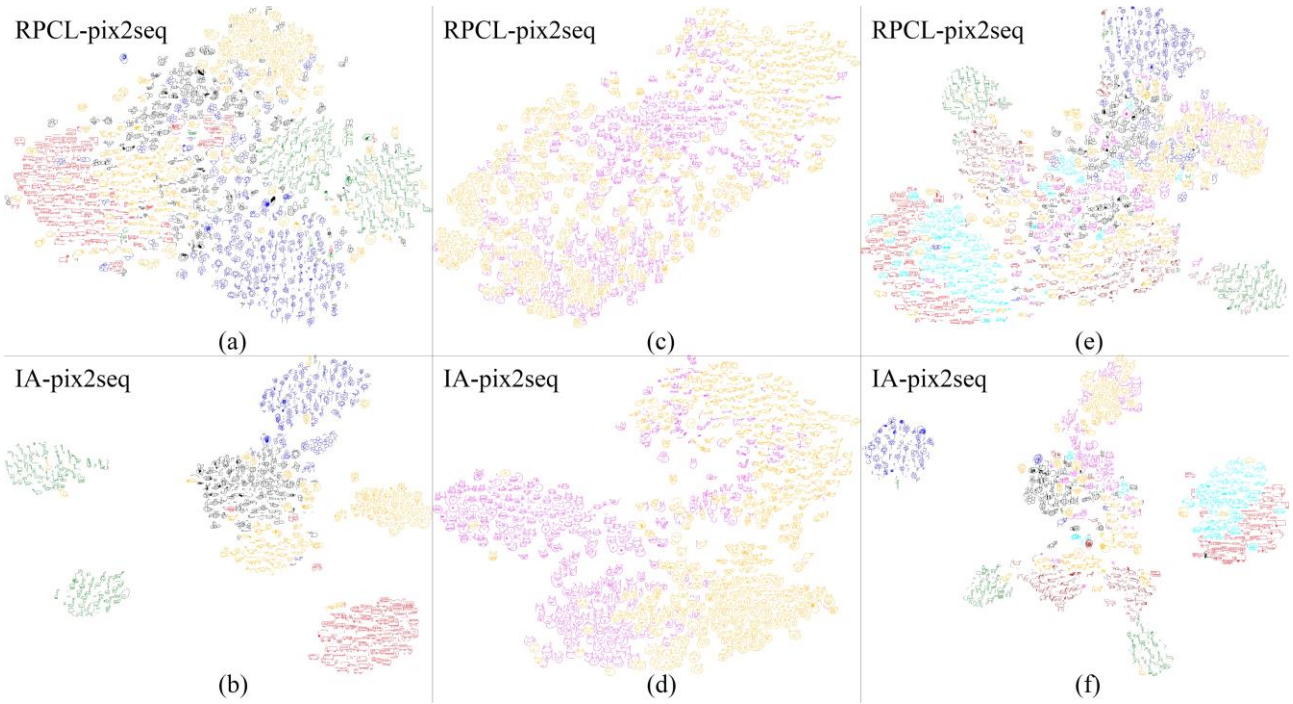

图 3　RPCL-pix2seq[1]和 IA-pix2seq 的编码空间可视化比较。图中的 3 列分别对应数据集 1-3 下的结果

图 3 使用 *t*-SNE[23]实现了高维编码空间的可视化。第 1-3 列分别对应数据集 1-3 下得到的编码空间；第 1-2 行分别对应了 RPCL-pix2seq 和 IA-pix2seq 的可视化结果。图中黑色、红色、青色、洋红色、蓝色、绿色、棕色和黄色的简笔画分别对应类别蜜蜂、公交车、汽车、猫、花、长颈鹿、马和猪。相比于 RPCL-pix2seq 构造的编码空间，IA-pix2seq 的编码空间呈现多个边界清晰、聚合紧凑的聚簇。如图 3(b)中，两个朝向的长颈鹿、公交车等编码区域间彼此分离，编码高斯间的重叠程度低于图 3(a)中 RPCL-pix2seq 的结果。如图 3(d)中，IA-pix2seq 的编码空间成功区分了猫头、猪头和带完整身体的动物形象，而图 3(c)中的 RPCL-pix2seq 无法做到。

为进一步量化比较 RPCL-pix2seq 和 IA-pix2seq 在 GMM 编码空间中高斯成分间的重叠程度，我们在高维编码空间中直接计算了两两高斯间的重叠率（Overlap Rate, OLR）[24]。如图 4，在 GMM 分布 $p(x)$ 中，*OLR* 度量高斯成分 G_1 和 G_2 间的重叠程度。*OLR* 的值域为 (0,1]，取值越大，重叠程度越严重。图 4 中的 x_{saddle} 和 $x_{\text{sub\_max}}$ 分别表示 G_1、G_2 间的鞍点和脊线上拥有较小概率密度的峰值所在位置。

我们同样根据线性判别分析（Linear Discriminant Analysis, LDA）[25]构造 R^{LDA}，如公式 (12)，以衡量高斯间的差异。

$$R^{\text{LDA}}_{G_i,G_j}=\frac{\left|\text{mean}(\boldsymbol{Y}_i)-\text{mean}(\boldsymbol{Y}_j)\right|_2^2}{\left|\text{std}(\boldsymbol{Y}_i)\right|_2^2+\left|\text{std}(\boldsymbol{Y}_j)\right|_2^2}. \tag{12}$$

式中 $\boldsymbol{Y}_i$ 表示测试集中被指派给编码高斯 G_i 的所有简笔画编码集合，mean(·)、std(·) 和 $|\cdot|_2$ 分别表示均值、标准差和 L2 范数。编码空间中高斯间的 *OLR* 越小、R^{LDA} 越大，表明两高斯的重叠程度越低、对应简笔画模式表达的差异越大，相应简笔画可控生成性能应越强。

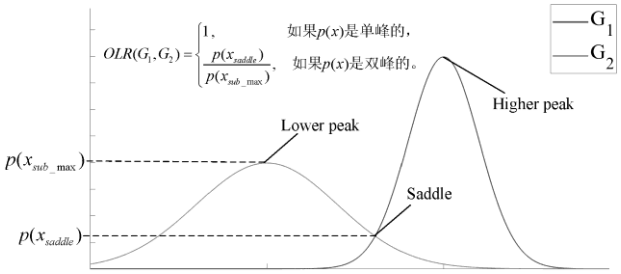

图 4　重叠率 *OLR*[24]的定义

表 4 给出了 RPCL-pix2seq 和 IA-pix2seq 所构造编码高斯间的重叠程度比较。IA-pix2seq 在三个数据集上的 *OLR* 均低于 RPCL-pix2seq，而在 R^{LDA} 上表现均优于 RPCL-pix2seq。这表明 RPCL-pix2seq 的编码高斯间的差异性小于本文提出的 IA-pix2seq。这是因为得益于 BYY 和谐学习，IA-pix2seq 限制了从简笔画到编码映射的信息传输量，缩小了每个简笔画在编码空间中的投影范围，鼓励 GMM 的编码指派逼近独热形式，以提高 GMM 各高斯的内聚度。编码空间中，不同类别和风格简笔画投影区域间的边界被拓宽，以保证 IA-pix2seq 对简笔画特征的区分性，使得 IA-pix2seq 获得了优于 RPCL-pix2seq 的可控生成性能，如表 1-表 3 所示。

SketchHealer 和 IA-pix2seq+的实验结果比较，进一步验证了使用 BYY 和谐学习算法的优越性。在保证编、解码器网络结构完全一致的条件下，IA-pix2seq+仍能保证编码高斯间较低的重叠程度和较高的差异性，如表 4。相应地，IA-pix2seq+较 SketchHealer 在三个数据集的可控生成性能上均有了进一步提升，如表 1-表 3。这同样反应了本文提出的 IA-pix2seq 在不同简笔画输入格式、不同编、解码器结构的适应和推广能力。

表 4 编码空间中高斯间的重叠程度（均值±标准差）比较。*OLR*[24]和 R^{LDA} 分别表示高斯重叠率和基于 LDA [25]的判别系数。所有指标均直接计算自高维空间中的编码

数据集	方法	*OLR*↓	R^{LDA} ↑
数据集 1	RPCL-pix2seq	0.3213±0.3250	0.1826±0.0700
	IA-pix2seq	0.0891±0.1045	1.6712±0.5356
	IA-pix2seq+	0.0405±0.1739	0.5920±0.1778
数据集 2	RPCL-pix2seq	0.4382±0.3059	0.3730±0.1389
	IA-pix2seq	0.3785±0.2137	0.7330±0.1247
	IA-pix2seq+	0.0804±0.1056	0.6034±0.0956
数据集 3	RPCL-pix2seq	0.2993±0.3195	0.5491±0.6709
	IA-pix2seq	0.1284±0.1533	1.3916±0.9351
	IA-pix2seq+	0.0485±0.1770	1.4791±0.1961

4.3 消融实验

IA-pix2seq 在简笔画生成模型的顶层、底层均引入了 BYY 和谐学习（HL），是与目前已有模型的主要区别。当 HL 被完全移除，IA-pix2seq 退化为最大似然（ML）学习。本节通过消融实验展示这两部分对性能的贡献。我们通过固定公式(8)中的 $\delta_{ik}=0$ 来去掉顶层的 HL，使顶层学习退化为 ML；通过将公式(2)中的 $H(p\|p)$ 项去掉，使底层退化为 ML。表 5-表 7 分别给出了三个数据集下的定量表现。

当 HL 仅作用于 IA-pix2seq 顶层时，指派给同一高斯的编码仍能紧凑地投影。在数据集 1 下的平均 *OLR* 仅为 0.1050，与表 4 中两层均使用 HL 的 IA-pix2seq 的平均 *OLR*=0.0891 接近。因此，两者在编码聚类的表现也较为接近，对应表 5 中 *Acc* 的 92.94%和 93.31%。而在面对简笔画特征提取更为困难的数据集 2-3 时，底层的 ML 无法令各简笔画 $\boldsymbol{x}_i$ 在编码空间中的投影区域 $p_\phi(\boldsymbol{y}|\boldsymbol{x}_i)$ 同步收缩，以适应高度紧凑的编码高斯 $q_{\boldsymbol{\theta}}(\boldsymbol{y}|k)$。这导致不同简笔画 $\boldsymbol{x}_i$、$\boldsymbol{x}_j$ 的投影区域 $p_\phi(\boldsymbol{y}|\boldsymbol{x}_i)$、$p_\phi(\boldsymbol{y}|\boldsymbol{x}_j)$ 严重重叠，影响了被投影至同一编码高斯内相同类别、风格简笔画间的检索表现，如表 5-表 7 中 *Ret* 所示。

当 HL 仅作用于 IA-pix2seq 底层时，顶层使用 EM 算法学习 GMM 结构，失去了自动模型选择功能。因此，初始化时引入的冗余编码高斯破坏了简笔画特征与编码高斯间的一一对应关系。相同类别和风格的简笔画被分散地投影到多个不同高斯内。因此，其在 *Ret* 上的表现明显低于较两层均使用 HL 的 IA-pix2eq。同时，以数据集 1 为例，“顶层 ML、底层 HL”的平均 *OLR* 为 0.1801，大于表 4 中 IA-pix2eq 的 0.0891，略优于 RPCL-pix2seq 的 0.3213。对比表 5-表 7 中两层均使用 ML 的结果，受 HL 作用的底层一定程度上紧凑了各简笔画 $\boldsymbol{x}_i$ 对应投影区域 $p_\phi(\boldsymbol{y}|\boldsymbol{x}_i)$ 的分布，一定程度上提高了 *Rec* 和 *Ret* 表现。

表 5 数据集 1 下和谐学习 HL 分别作用于与 IA-pix2seq 的顶层、底层后，简笔画可控生成性能比较（%）。未使用 HL 的 IA-pix2seq 结构则使用最大似然学习 ML

学习策略		*Rec*↑	*Ret*↑			*Acc*↑
顶层	底层		Top-1	Top-10	Top-50	
ML	ML	83.45	3.66	13.61	27.22	76.00
ML	HL	89.86	8.10	29.51	54.55	74.83
HL	ML	92.02	0.90	6.17	20.08	92.94
HL	HL	**94.75**	**24.34**	**48.08**	**67.36**	**93.31**

表 6 数据集 2 下和谐学习 HL 分别作用于与 IA-pix2seq 的顶层、底层后，简笔画可控生成性能比较（%）

学习策略		*Rec*↑	*Ret*↑			*Acc*↑
顶层	底层		Top-1	Top-10	Top-50	
ML	ML	71.00	1.52	7.90	19.06	62.16
ML	HL	73.50	6.00	19.00	36.42	61.96
HL	ML	79.68	4.70	20.06	41.64	69.78
HL	HL	**82.74**	**24.38**	**50.18**	**68.54**	**81.12**

表 7 数据集 3 下和谐学习 HL 分别作用于与 IA-pix2seq 的顶层、底层后，简笔画可控生成性能比较（%）

学习策略		*Rec*↑	*Ret*↑			*Acc*↑
顶层	底层		Top-1	Top-10	Top-50	
ML	ML	60.37	0.26	1.92	7.09	61.80
ML	HL	74.99	6.24	24.02	46.19	55.75
HL	ML	57.70	0.31	22.54	49.09	53.83
HL	HL	**83.99**	**37.28**	**58.09**	**83.53**	**76.50**

总而言之，和谐学习需同时作用于 IA-pix2seq 的顶层和底层，才能更好地保持编码空间中简笔画的投影区域大小和编码高斯内聚度间的平衡，将简笔画可控生成性能保持在高水平。

5 IA-pix2seq 应用于简笔画可控生成

本文提出的 IA-pix2seq 在编码空间中形成了诸多彼此分离而又内部紧凑的高斯区域，分别对应真实空间中不同简笔画类别和风格特征。这允许 IA-pix2seq 通过内在表达（即编码数值）或是外部输入（即简笔画图片）的约束，操控简笔画生成。

5.1 编辑内在表达：在编码空间中插值

IA-pix2seq 将具有相似特征模式的简笔画编码更集中地投影到编码空间的相应高斯中心。高斯间更低的重叠率，使得编码空间中出现大面积未被测试集简笔画覆盖的区域。若上述区域中的编码能被 IA-pix2seq 解码为结构合理、形态清晰的简笔画，说明 IA-pix2seq 具有良好的简笔画可控生成泛化能力。图 5 展示了 IA-pix2seq 在数据集 1 上插值生成

的简笔画，简笔画的不同灰度表示在两个方向插值时的系数。图 5 插值结果的四个端点分别对应四个高斯的中心编码，位于中央部分简笔画的编码来自未被测试集覆盖的编码区域，以验证 IA-pix2seq 在简笔画生成上的泛化能力。

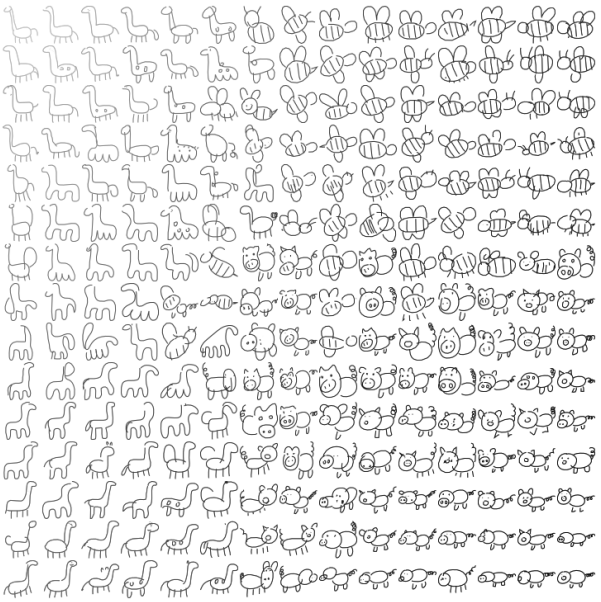

图 5　在 IA-pix2seq 的编码空间中插值。插值的四个端点分别对应编码空间中四个高斯的中心编码。简笔画的不同灰度表示在两个方向插值时的系数。

分析图 5 的插值结果：从左上角水平向右，长颈鹿的脑袋逐渐变大，脖子变粗变短，四肢收缩，身体由扁平转为圆润，而后出现斑纹，乃至蜜蜂翅膀的成型；从左上角至右下角，长颈鹿的身体逐渐丰满，头部显著变大，而后出现了显著的猪鼻子特征，完成从长颈鹿到猪的过渡。进而猪的身体扁平化，四肢逐渐出现，最终形成右下角的带完整身体的猪。位于图 5 中央部分简笔画对应的编码，虽然来自未被测试集样本覆盖的编码区域，但生成结果上并未出现过渡突兀、不合理的形象。这表明 IA-pix2seq 构建的编码空间相对平滑，并没有因为 GMM 高斯的高内聚，造成生成简笔画形象上的突变。测试集简笔画聚集在编码空间中高斯中心区域，为保证其类别的可识别性，相应的简笔画往往呈现相对规则、“保守”的形象；而高斯间编码区域对应的简笔画，则对应设计更为灵活、特征更为丰富的形象。相比于 sketch-pix2seq 等将测试集样本更均匀覆盖编码空间的方法，IA-pix2seq 更容易生成新颖、有创意的简笔画。

图 6 给出了 IA-pix2seq 在编码空间沿箭头方向插值时生成的创意简笔画。箭头上方的插值结果保留了箭头两端简笔画的部分特征，形成了有趣的新形象。如第 1 行第 1 列，在不同朝向的长颈鹿间插值，生成了“回眸”的长颈鹿，这在数据集中未曾出现；第 2 行第 1 列，插值结果同时保留了猫头和花茎。

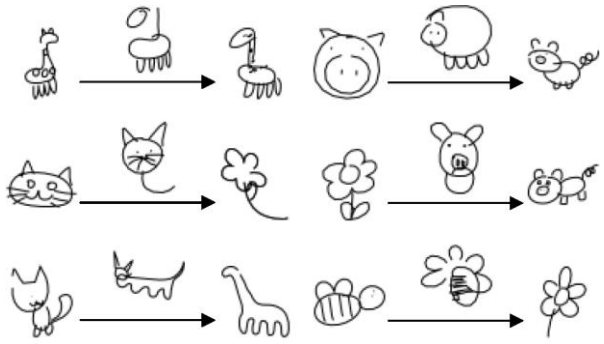

图 6　IA-pix2seq 在编码空间沿箭头方向插值时生成的创意简笔画

5.2　调整外部输入：简笔画补全

对于外部的简笔画输入，IA-pix2seq 根据其类别和风格，经内部编码后生成与之特征相似的简笔画“重建”。而当输入不再清晰完整，IA-pix2seq 只能从信息缺失的简笔画中获得部分信息导向。本节实验将展示各方法对残缺简笔画的还原、补全能力，如图 7。

图 7 中前 3 行分别展示了原始简笔画、含空白干扰像素的蒙版（白色区域为被抹除信息的位置）和信息缺失的简笔画输入。不规则形状的蒙版采用文献[26]中的方法随机生成。图中第 4 至 8 行展示了各方法的简笔画补全表现。sketch-rnn 和 SketchHealer 的输入不为简笔画图片，未参与本实验比较。图 7 中，还原后无法被准确识别为原始输入类别的简笔画被虚线框标记；生成质量较差的简笔画被点划线框标记。对于仍保留了关键特征的简笔画，如第 4 列中保留了车轮和部分车窗的公交车，第 8 列中保留了头部、四肢和身体朝向的长颈鹿，各方法基本可以还原原始简笔画的类别和风格。当简笔画的部分特征缺失时，IA-pix2seq 还原了原始简笔画的更多细节，如第 1 列中生成的蜜蜂与输入均没有翅膀，第 12 列中生成的猫头两侧分别对应一根和二根胡须。IA-pix2seq 所构造的紧凑、鲁棒的编码空间，保证在面对部分非关键信息缺失的简笔画输入时，仍能识别并保留更多的简笔画细节。而当简笔画的关键特征明显缺失，如第 2 列中几乎没有斑纹的蜜蜂，简笔画的识别难度剧增，所有方法均无法在还原简笔画中保持同原始类别的一致性。

6　结　论

本文提出了用于简笔画可控生成的深度双向模型 IA-pix2seq。模型分为两层，顶层采用 GMM 分布描述简笔画的编码空间，底层是单编码-双解码结构，即由一个 CNN 编码器将简笔画映射到编码空间、一个 CNN 解码器重建原简笔画、另一个 RNN

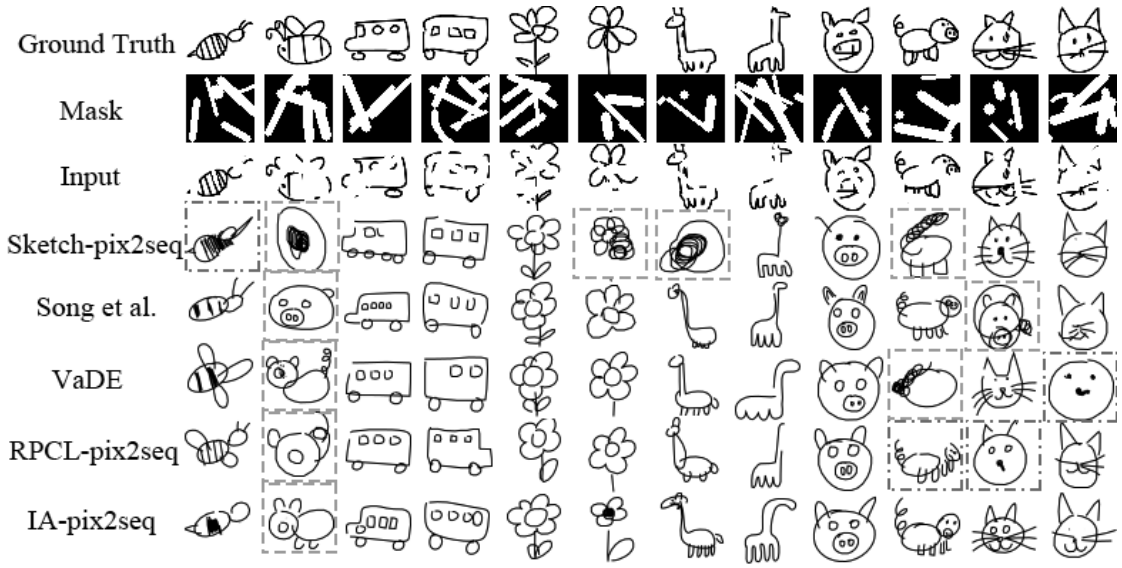

图 7 基于不完整的简笔画输入，各方法还原原始简笔画的表现。图中虚线框标记了还原生成的简笔画无法被准确识别为原始输入类别；点划线框标记了生成质量较差的简笔画。

解码器作可控生成。双向互逆映射在和谐学习原理指导下，以最默契的方式达到最大共识，在学习算法中引入了高效的竞争进入和惩罚原理机制，不仅使得编码空间 GMM 的高斯数可以自动确定，而且形成诸多彼此显著分离而又内部紧密集中的高斯成分编码区域，构建了从外部简笔画数据特征与内部编码概念的一一对应关系。大量实验结果表明，IA-pix2seq 相比已有方法有效地提高了简笔画可控生成性能。对于编辑后的编码或是修改后的简笔画输入，IA-pix2seq 均能在生成简笔画上保留以上输入约束的细节特征，并生成创意新颖、结构合理的简笔画。

参 考 文 献

[1] Zang S, Tu S, Xu L. Controllable stroke-based sketch synthesis from a self-organized latent space. Neural Networks, 2021, 137: 138-150

[2] Chen X, Duan Y, Houthooft R, et al. InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets//Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona, Spain, 2016: 2180-2188

[3] Gatys L A, Ecker A S, Bethge M. Image style transfer using convolutional neural networks//Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 2414-2423

[4] Ha D, Eck D. A neural representation of sketch drawings//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada, 2018

[5] Sangkloy P, Burnell N, Ham C, et al. The Sketchy database: Learning to retrieve badly drawn bunnies. ACM Transactions on Graphics, 2016, 35(4): 1-12

[6] Eitz M, Hays J, Alexa M. How do humans sketch objects? ACM Transactions on Graphics, 2012, 31(4): 1-10

[7] Xu L. Learning deep IA bidirectional intelligence. Frontiers of Information Technology & Electronic Engineering, 2020, 21(4): 558-562

[8] Xu L. An overview and perspectives on bidirectional intelligence: Lmser duality, double IA harmony, and causal computation. IEEE/CAA Journal of Automatica Sinica, 2019, 6(4): 865-893

[9] Xu L. Bayesian-Kullback coupled Ying-Yang machines: unified learnings and new results on vector quantization//Proceedings of 1995 International Conference on Neural Information Processing. Beijing, China, 1995: 977-988

[10] Xu L. A unified learning scheme: Bayesian-Kullback Ying-Yang machine//Proceedings of the 8th Advances in Neural Information Processing Systems. Denver, USA, 1995: 444-450

[11] Chen Y, Tu S, Yi Y, Xu L. Sketch-pix2seq: a model to generate sketches of multiple categories. arXiv preprint arXiv:1709.04121, 2017

[12] Xu L, Krzyzak A, Oja E. Rival penalized competitive learning for clustering analysis, RBF net, and curve detection. IEEE Transactions on Neural Networks, 1993, 4(4): 636-649

[13] Jordan M, Xu L. Convergence results for the EM approach to mixtures of experts architectures. Neural Networks, 1995, 8(9): 1409-1431

[14] Redner R, Walker H. Mixture densities, maximum likelihood and the EM algorithm. SIAM Review, 1984, 26(2): 195-239

[15] Jiang Z, Zheng Y, Tan H, et al. Variational deep embedding: An unsupervised and generative approach to clustering//Proceedings of the 26th International Joint Conference on Artificial Intelligence. Melbourne, Australia, 2017: 1965-1972

[16] Ha D, Dai A, Le Q. Hypernetworks//Proceedings of the 5th International Conference on Learning Representations. Toulon, France,

2017

[17] Kingma, D P, Welling M. Auto-encoding variational Bayes. arXiv preprint arXiv:1312.6114, 2013

[18] Xu L. Bayesian Ying-Yang system, best harmony learning, and five action circling. Frontiers of Electrical and Electronic Engineering in China, 2010, 5(3): 281-328

[19] Shi L, Tu S, Xu L. Learning Gaussian mixture with automatic model selection: A comparative study on three Bayesian related approaches. Frontiers of Electrical and Electronic Engineering in China, 2011, 6(2): 215-244

[20] Song J, Pang K, Song Y, et al. Learning to sketch with shortcut cycle consistency//Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 801-810

[21] Su G, Qi Y, Pang K, et al. SketchHealer: A graph-to-sequence network for recreating partial human sketches//Proceedings of the 31st British Machine Vision Conference. Virtual Event, UK, 2020

[22] Yu Q, Yang Y, Song Y, et al. Sketch-a-Net that beats humans. arXiv preprint arXiv:1501.07873, 2015

[23] Maaten L, Hinton G. Visualizing data using t-SNE. Journal of Machine Learning Research, 2008, 9(11): 2579-2605

[24] Sun H, Wang S. Measuring the component overlapping in the Gaussian mixture model. Data Mining and Knowledge Discovery, 2011, 23(3): 479-502

[25] Fisher R. The use of multiple measurements in taxonomic problems. Annals of Eugenics, 1936, 7(2): 179-188

[26] Zheng C, Cham T, Cai J. Pluralistic image completion//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019: 1438-1447

ZANG Si-Cong, Ph. D. candidate. His main research interest includes machine learning.

TU Shi-Kui, Ph. D., tenure-track associate professor. His main research interests include machine learning and bioinformatics.

Xu Lei, Ph. D., professor. His main research interests include machine learning, causal computation, bidirectional intelligence, AlphaGo-inspired system, computational precision medicine and computational finance.

Background

In the human history, drawing free-hand sketches is a simple way for communication and expression. Sketches are always abstract, lack-of-details and variant, but can still convey vivid messages and emotions as the pixel-formed images. Moreover, free-hand sketches are more flexible and practical to convey conceptually hybrid information with structurally unique designs. The first step for freely building such sketch designs is making the sketch synthesis controllable.

To obtain the controllable sketch synthesis process, the recent studies target to build a latent space, preserving the similarity of structural patterns from the observed sketch data to the latent codes. Such a latent space is regarded as a Voronoi tessellation, where each latent region is assigned a unique concept or a pattern, representing a specific sketch category and style. Thus, it is practical to locate a specific latent code to synthesize a sketch with the expected patterns. A group of studies utilized different forms of sketches, such as the sequential form, the pixel form or both, expecting to extract more sketch patterns to assist the latent space self-organization. Another group focused on the structure of the latent space, e.g., aiming to self-organize a single Gaussian distributed, a uniform distributed or a Gaussian mixture model (GMM) distributed latent space. These latent structures help fit the real sketch data distribution properly, encouraging the sketch synthesis to be controllable.

The Gaussian regions from the GMM latent space build by the recent models are heavily overlapped, which reduces the controllable synthesis performance. This paper present IA-pix2seq which is guided by the Bayesian Ying-Yang (BYY) harmony learning algorithm. BYY harmony learning seeks a best matching between encoding and decoding subsystems with a most tacit manner by minimizing the information transferred from sketches to latent codes. Correspondingly, IA-pix2seq not only centralizes the latent variables within a latent Gaussian component but also squeezes the latent territory for each sketch sample. The experimental results show the sketch synthesis process are more easily controlled and the generated sketches preserve more details of the input constraint which can be either an interpolated latent code or a masked sketch image.

This work was supported by National Science and Technology Innovation 2030 Major Project (2018AAA0100700) of the Ministry of Science and

Technology of China, and Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102).