第 38 卷　第 5 期
2023 年 10 月

天津科技大学学报
Journal of Tianjin University of Science & Technology

Vol. 38　No. 5
Oct. 2023

DOI:10.13364/j.issn.1672-6510.20230036

全相关约束下的变分层次自编码模型

陈亚瑞，胡世凯，徐肖阳，张　奇
（天津科技大学人工智能学院，天津 300457）

摘　要：基于深度学习的解耦表示学习可以通过数据生成的方式解耦数据内部多维度、多层次的潜在生成因素，并解释其内在规律，提高模型对数据的自主探索能力。传统基于结构化先验的解耦模型只能实现各个层次之间的解耦，不能实现层次内部的解耦，如变分层次自编码（variational ladder auto-encoders，VLAE）模型。本文提出全相关约束下的变分层次自编码（variational ladder auto-encoder based on total correlation，TC-VLAE）模型，该模型以变分层次自编码模型为基础，对多层次模型结构中的每一层都加入非结构化先验的全相关项作为正则化项，促进此层内部隐空间中各维度之间的相互独立，使模型实现层次内部的解耦，提高整个模型的解耦表示学习能力。在模型训练时采用渐进式训练方式优化模型训练，充分发挥多层次模型结构的优势。本文最后在常用解耦数据集 3Dshapes 数据集、3Dchairs 数据集、CelebA 人脸数据集和 dSprites 数据集上设计对比实验，验证了 TC-VLAE 模型在解耦表示学习方面有明显的优势。

关键词：解耦表示学习；变分自编码器；概率生成模型；结构化先验；非结构化先验
中图分类号：TP181　　**文献标志码**：A　　**文章编号**：1672-6510（2023）05-0064-10

Variable Ladder Auto-Encoder Based on Total Correlation

CHEN Yarui，HU Shikai，XU Xiaoyang，ZHANG Qi
（College of Artificial Intelligence，Tianjin University of Science & Technology，Tianjin 300457，China）

Abstract：Decoupling presentation learning based on deep learning can decouple multi-dimensional and multi-level potential generation factors within data by means of data generation，and explain their internal rules to improve the model's ability to explore data independently. However，the traditional decoupling model based on structured priori can only realize the decoupling between all levels，but cannot realize the decoupling within the levels，such as variational ladder auto-encoders（VLAE）. In this article，we first propose a variational ladder auto-encoder based on total correlation（TC-VLAE），which is based on the variational ladder auto-encoder based on total correlation. Then，the full correlation terms of unstructured priors are added to each layer of the multi-level model structure as regularization terms to promote the independence of each dimension in the hidden space inside this layer，so that the model can realize the decoupling within the hierarchy and improve the decoupling representation learning ability of the whole model. In model training，a progressive training method is adopted to optimize model training and give full play to the advantages of multi-level model structure. Finally，comparative tests are designed on the commonly used decoupling data sets 3Dshapes data set，3Dchairs data set，CelebA face data set and dSprites data set，and it is verified that TC-VLAE model has obvious advantages in decoupling representation learning.

Key words：disentanglement learning；variational autoencoder；probabilistic generative model；structured prior；unstructured prior

深度学习的发展需要探索数据特征并解释其内在规律。在深度学习领域对学习揭露数据特征内在规律的可解释性表示的研究并不多。这种可解释性表示可以让人们观察到原始数据的特征表示并能够捕捉与最终任务相关的潜在（抽象或高级）生成因素，同时忽略不合理或无用的因素。这种可解释性表示

收稿日期：2023-02-23；**修回日期**：2023-05-05
基金项目：天津科技大学青年教师资助计划项目（2017LG10）
作者简介：陈亚瑞（1982—），女，河北邢台人，副教授，yrchen@tust.edu.cn

在机器学习算法和深度学习算法的研究中发挥着重要作用[1]。这种表示不仅对监督学习和强化学习等标准下游任务有用，而且对那些人类擅长而机器不擅长的任务也有非常大的帮助，比如迁移学习和零样本学习等[2]。

解耦表示学习是深度学习领域中学习数据特征的可解释性表示的探索研究。假设数据由多个维度且数量固定的独立生成因子生成，其中一个维度的变化只对应一个数据的变化[3]。解耦表示学习旨在按照人类能够理解的方式从真实数据中对具有明确物理含义的生成因子（如类别、位置、外观、纹理等）进行解耦，并给出其对应的独立表示。解耦表示学习不仅在探索数据的可解释性方面存在显著优势，而且应用场景广阔，比如图像编辑、图像生成和 3D 建模等。解耦表示学习逐渐成为深度学习领域的重要研究方向并引起国内外众多学者的广泛关注[4]。

早期的解耦表示学习研究可以追溯到独立成分分析（independent component algorithm，ICA）。该方法假设信号是由多种独立成分线性叠加而成，线性解耦方法的应用范围和深度都极为有限[4]。随着深度学习研究的不断深入，基于神经网络的深度生成模型在数据（尤其是图像）解耦表示学习方面显示出巨大的前景，比如在推荐领域中的挖掘用户多样偏好。在解耦表示学习的发展初期，研究者们大多以监督学习或者半监督学习的方式在带有标签或者少量标签的数据集上进行解耦表示学习的研究。由于数据量的不断扩大导致人工标注的成本不断增加，同时人工标签还存在可能与实际数据不一致或遗漏人类难以识别的因素等缺点，因此研究者们越来越注重以无监督的方式进行解耦表示学习的研究。

随着相关研究的不断深入，越来越多无监督方式的解耦表示学习算法被提出，其中两种主流的研究思路是以生成对抗网络（generative adversarial nets，GAN）[5]和变分自编码（variational auto-encoder，VAE）模型[6]作为基础进行解耦研究。基于 GAN 的解耦表示学习模型有信息最大化生成对抗网络（information maximizing generative adversarial nets，Info-GAN）和信息蒸馏生成对抗网络（information-distillation generative adversarial network，ID-GAN）。GAN 模型只注重数据生成过程而缺乏推理过程，并且还存在模型坍塌的问题，这些都妨碍了 GAN 模型在解耦表示学习领域的发展。VAE 模型有完整的推理过程和生成过程，这为解耦表示学习的研究打下了良好的基础。

Higgins 等[9]在 2017 年提出了 β-VAE 模型，它是在 VAE 模型损失函数的 KL 项上添加一个额外的并且大于 1 的超参数 β，这可以使模型学习到具有统计独立性的隐变量，从而使模型具有一定的解耦表示学习能力。但是，大于 1 的超参数 β 会降低重构误差的权重，导致模型重建数据的能力较差，即生成图片的质量较低。Burgess 等[10]通过在训练过程中逐渐增加隐变量的信息容量，解决了 β-VAE 模型不能很好地平衡重建数据质量和解耦表示学习能力的问题。Chen 等[11]提出了 β-TCVAE（β-total correlation variational auto-encoder）模型，它将 VAE 模型的 KL 项分解为 3 项，其中的全相关项（total correlation，TC）是模型解耦能力大小的关键。β-TCVAE 模型通过一个大于 1 的超参数 β 加大 TC 项在模型训练中的权重，使模型的解耦表示学习能力得到大幅度提高。在计算 TC 项时，β-TCVAE 模型基于重要性采样的思想对批量样本进行加权采样，这种方式具有简单、高效且模型训练稳定的优点。

Kim 等[2]提出的 Factor-VAE 模型将 VAE 模型损失函数的 KL 项分解为两项，分别为数据与隐变量之间的互信息、隐变量的聚合后验分布和先验分布之间的 KL 散度。本课题组认为加重第二项的权重可以显著提升模型的解耦表示学习能力，但不同的是本研究并没有直接加重第二项的权重，而是在 VAE 模型损失函数不变的情况下，加入一项 TC 项作为约束隐变量的正则化项，这同样可以促使隐空间中各维度之间互相独立，从而激励模型学习更好地解耦表示学习能力。这类通过探索隐空间进行解耦表示学习研究的模型可以总结为基于非结构化先验的解耦表示学习模型。

对于解耦表示学习，Montero 等[12]提出了不同的思路，他们认为设计由人类认知过程启发的高度显示结构化的网络模型对解耦表示学习进行研究尤为重要。通过现实世界中许多自然数据本身所特有的成分分层特性，设计搭建了层次深度梯形网络模型，通过组合较低层的语义特征获得较高层的语义特征表示[4]，这可以总结为基于结构先验的解耦表示学习模型。典型的模型有 Sønderby 等[13]将层次深度梯形网络与变分自编码器结合，提出了层次变分自编码（ladder variational auto-encoders，LVAE）模型。与传统 VAE 模型不同，LVAE 模型提出推理与生成模型共享自顶向下的依赖结构，使模型的推理过程只用简单的先验分布，将优化过程变得更加容易。

LVAE 建立的层次网络结构有局限性：如果这些模型可以训练为最优，那么第一层的信息就足以重建数据分布，而第一层之上的层可以忽略。Zhao 等[14]提出变分层次自编码（variational ladder auto-encoders，VLAE）模型。该模型将不同层次的隐变量与具有不同表达能力（深度）的网络连接，鼓励模型在顶部放置高层次、抽象的特征（如身份特征等），在底部放置低层次、简单的特征（如边缘特征等）。这种模型设计使得越高层、越抽象的特征需要越复杂的网络捕获，在不需要特定先验知识的情况下，能够学习高度可解释的、解耦的层次特征[4]。

本文在基于 VLAE 模型框架下提出了全相关约束下的变分层次自编码（variational ladder auto-encoder based on total correlation，TC-VLAE）模型，该模型融合了基于非结构先验和结构先验两种方法。TC-VLAE 模型基于层次化的变分自编码模型，层次化的网络模型可以实现层次之间的解耦。在每一层的隐空间中都加入非结构先验的 TC 项作为正则化项，TC 项可以促进每一层隐空间中各个维度的隐变量之间相互独立，从而实现单个层内部的解耦。TC-VLAE 模型同时实现了层级之间和层级内部的解耦。TC-VLAE 模型在训练时使用渐进式的训练算法训练模型，如 Karras 等[15]和 Wang 等[16]的工作。渐进式的训练算法通过分步训练，将多层隐变量逐步加入模型训练中，充分发挥层次化模型的优势，有利于模型的稳定训练。同时，使用随机梯度下降算法求解模型参数。为了验证模型的有效性，分别在 3Dshapes 数据集、3Dchairs 数据集、CelebA 数据集和 dSprites 数据集上设计实验，验证 TC-VLAE 模型具有更好的解耦表示学习能力。

1　变分层次自编码模型

变分层次自编码模型（VLAE）是在变分自编码模型的基础上引入层次化结构先验，将隐变量分解到不同层次上，利用层次结构进行解耦表示学习。

VLAE 将隐变量 $\boldsymbol{z}$ 分解为 L 部分 $\boldsymbol{z}=\{\boldsymbol{z}_1,\boldsymbol{z}_2,\cdots,\boldsymbol{z}_L\}$，每部分隐变量构成模型的一层，其具体网络结构如图 1 所示。图 1 是一个有 L 层隐变量的 VLAE 模型的网络架构，其中 $\boldsymbol{x}$ 代表可观测数据，$\boldsymbol{z}$ 表示隐变量，菱形表示确定性的节点，$\boldsymbol{h}_l$ 和 $\tilde{\boldsymbol{z}}_l$ 分别表示相应层深度神经网络的确定性输出。

模型的推理过程为

$$\boldsymbol{h}_l = g_l(\boldsymbol{h}_{l-1}) \tag{1}$$

$$\boldsymbol{z}_l \sim N(\mu_l(\boldsymbol{h}_l),\sigma_l(\boldsymbol{h}_l)) \tag{2}$$

其中：$l=1,\cdots,L$，表示 VLAE 模型的隐变量层次，g_l、μ_l、σ_l 表示每层所用的神经网络，令 $\boldsymbol{h}_0 \equiv \boldsymbol{x}$。

模型的生成过程为

$$\tilde{\boldsymbol{z}}_L = f_L(\boldsymbol{z}_L) \tag{3}$$

$$\tilde{\boldsymbol{z}}_l = u_l([\tilde{\boldsymbol{z}}_{l+1};v_l(\boldsymbol{z}_l)]) \tag{4}$$

其中：f_L 作为解码器是参数化的神经网络，[；]表示两个向量的连接，v_l、u_l 表示神经网络。

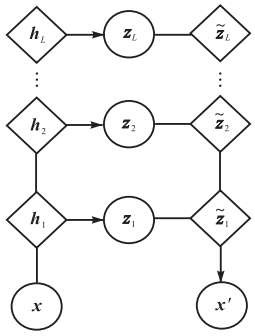


图 1　VLAE 网络结构

Fig. 1　VLAE network structure

模型的推理过程：模型的输入数据 $\boldsymbol{x}$ 通过神经网络 g_1 得到 $\boldsymbol{h}_1$，如式（1）所示。$\boldsymbol{h}_1$ 一方面向左通过神经网络 μ_1、σ_1 得到第 1 层的隐变量 $\boldsymbol{z}_1$，如式（2）所示；另一方面向上通过神经网络 g_2 得到 $\boldsymbol{h}_2$。在第 2 层 $\boldsymbol{h}_2$ 执行与第 1 层相同的操作，得到第 2 层的隐变量 $\boldsymbol{z}_2$ 和 $\boldsymbol{h}_3$。以此类推直到第 L 层，得到 L 层的隐变量 $\boldsymbol{z}_1,\boldsymbol{z}_2,\cdots,\boldsymbol{z}_L$。

模型的生成过程：从网络结构的最顶层第 L 层开始，首先是第 L 层的隐变量 $\boldsymbol{z}_L$ 经过采样等操作再通过神经网络 f_L 得到 $\tilde{\boldsymbol{z}}_L$，如式（3）所示。第 L–1 层的隐变量 $\boldsymbol{z}_{L-1}$ 也经过采样等操作再通过神经网络 v_{L-1}，得到的结果 $v_{L-1}(\boldsymbol{z}_{L-1})$ 与 $\tilde{\boldsymbol{z}}_L$ 相连接并再次通过神经网络 u_{L-1} 得到 $\boldsymbol{z}_{L-1}$，如式（4）所示，直到求出 $\tilde{\boldsymbol{z}}_1$。此时的 $\tilde{\boldsymbol{z}}_1$ 就相当于 VAE 中的隐变量，对其进行采样、重参数化等操作即可重构图像数据 $\boldsymbol{x}$。VLAE 模型在训练时使用传统 VAE 模型的损失函数。

2　本文模型

TC-VLAE 模型在变分层次自编码模型基础上，通过在每层隐变量增加 TC 正则化先验约束，促使每一层的隐变量都具有解耦表示学习能力，然后通过网络结构将所有隐变量的解耦表示学习能力叠加。同时，采用渐进式方式进行模型训练，提升整个模型的

解耦效果。

2.1　模型结构

TC-VLAE 模型以多层次的变分层次自编码模型的网络结构为基础，并在每一层的隐变量中引入 TC 正则化先验约束。令 $\boldsymbol{x}\in\boldsymbol{R}^{\mathrm{D}}$ 表示观测向量，$\boldsymbol{z}\in\boldsymbol{R}^{\mathrm{M}}$ 表示低维连续隐向量。假设图像数据 $\boldsymbol{x}$ 是由隐变量 $\boldsymbol{z}$ 生成，采用具有 L 层的多层次网络结构将隐变量分为 L 层，即 $\boldsymbol{z}=\{\boldsymbol{z}_1,\boldsymbol{z}_2,\cdots,\boldsymbol{z}_L\}$，$\tilde{\boldsymbol{z}}_l$ 是经神经网络得到的确定性的节点，$l=1,2,\cdots,L$。模型定义为

$$p(\boldsymbol{x},\boldsymbol{z})=p(\boldsymbol{x}\mid\tilde{\boldsymbol{z}}_1)\prod_{l=1}^{L}P(\boldsymbol{z}_l) \tag{5}$$

即

$$\begin{aligned}&\tilde{\boldsymbol{z}}_1=f_1(\tilde{\boldsymbol{z}}_2,\boldsymbol{z}_1)\\&\tilde{\boldsymbol{z}}_2=f_2(\tilde{\boldsymbol{z}}_3,\boldsymbol{z}_2)\\&\quad\cdots\\&\tilde{\boldsymbol{z}}_{L\text{-}1}=f_{L-1}(\tilde{\boldsymbol{z}}_L,\boldsymbol{z}_{L-1})\\&\tilde{\boldsymbol{z}}_L=f_L(\boldsymbol{z}_L)\end{aligned}$$

其中：$p(\boldsymbol{z}_l)=N(\boldsymbol{z}_l;0,\boldsymbol{I})$ 表示隐变量的先验概率分布，$l=1,2,\cdots,L$，$\boldsymbol{I}$ 表示单位矩阵，$f_1,f_2,\cdots,f_L$ 为神经网络；$p(\boldsymbol{x}\mid\tilde{\boldsymbol{z}}_1)=N(\boldsymbol{x};\mu,\sigma^2\boldsymbol{I})$ 表示条件概率分布。

图像数据通过较少的卷积神经网络得到的隐变量只包含比较低层次的粗粒度特征信息，比如图片的背景、物体的大小等特征信息。图像数据通过较多的卷积神经网络得到的隐变量包含较高层次的细粒度特征信息，比如物体的颜色、形状和纹理等特征信息。VLAE 模型的多层次网络结构可以在不同层次捕捉不同的数据特征信息。

在此模型中每一层的隐变量都引入一个 TC 项作为约束项[11]，促使模型具有更好的解耦表示学习能力。TC 项为

$$\mathrm{KL}[q(\boldsymbol{z})\|\prod_i q(\boldsymbol{z}_i)] \tag{6}$$

其中：$q(\boldsymbol{z})$ 是聚合后验，也就是隐空间；i 是隐空间中隐变量的维度。KL 散度代表了两个分布之间的距离，上式的 TC 项越小，聚合后验分布与隐空间中各维隐变量分布乘积之间的距离就越小，两个分布就越相似，隐空间中各维隐变量之间就越独立，模型就具有了更强的解耦表示学习能力。

完整模型网络架构如图 2 所示，其中 $\boldsymbol{x}$ 为可观测数据，$\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_L$ 和 $\tilde{\boldsymbol{z}}_1,\tilde{\boldsymbol{z}}_2,\cdots,\tilde{\boldsymbol{z}}_L$ 为确定性的节点，$\boldsymbol{z}_1,\boldsymbol{z}_2,\cdots,\boldsymbol{z}_L$ 为隐变量；红色虚线框是 TC 项，梯形框是神经网络。

在模型推理过程中，数据 $\boldsymbol{x}$ 经过卷积神经网络得到 $\boldsymbol{h}_1$，$\boldsymbol{h}_1$ 向下通过神经网络得到隐变量 $\boldsymbol{z}_1$；$\boldsymbol{h}_1$ 再向右经过卷积神经网络得到 $\boldsymbol{h}_2$，$\boldsymbol{h}_2$ 再通过神经网络计算得到 $\boldsymbol{z}_2$；以此类推到第 L 层。不同层次的隐变量包含了不同层次的数据信息，较低的层次包含了粗粒度的数据特征信息，较高的层次包含了细粒度的数据特征信息。

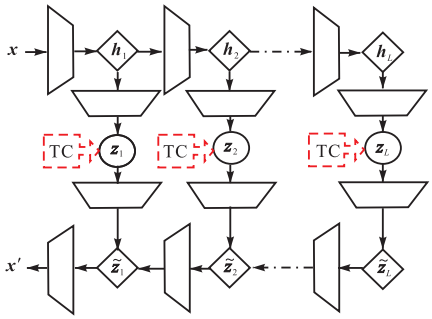


图 2　TC-VLAE 网络结构

Fig. 2　TC-VLAE network structure

在模型生成过程中，第 L 层的隐变量通过神经网络得到 $\tilde{\boldsymbol{z}}_L$，第 L–1 层的隐变量通过神经网络得到的结果与 $\tilde{\boldsymbol{z}}_L$ 连接并再次通过神经网络得到 $\tilde{\boldsymbol{z}}_{L-1}$，以此类推得到 $\tilde{\boldsymbol{z}}_1$，再经过重参数化等操作生成数据 $\boldsymbol{x}$。

这样生成过程中每一层的隐变量会直接影响下一层 $\tilde{\boldsymbol{z}}$ 的生成，而直接参与图片重构的最底层 $\tilde{\boldsymbol{z}}_1$ 会包含所有隐变量的信息，这使所有层的隐变量都会影响最终模型的生成结果。在多层次的网络模型中，只有最底层的 $\tilde{\boldsymbol{z}}_1$ 直接参与数据重构，而每一层的隐变量 $\boldsymbol{z}_l$ 只参与生成本层的 $\tilde{\boldsymbol{z}}$。

本文在 L 层隐变量中的每一层都引入 TC 项作为正则化项（如图 2 中红色虚线框所示）约束隐变量。这使每一个层次的隐变量都具有一定的解耦表示学习能力，低层次的隐变量具有解耦粗粒度特征的能力，高层次的隐变量具有解耦细粒度特征的能力，再通过模型的生成过程将所有隐变量的信息包含在一起，生成新的数据，最终使模型既能够解耦粗粒度特征，又能够解耦细粒度特征。

2.2　模型损失函数

为了避免直接修改传统模型损失函数对模型生成质量的影响，本文在保持传统模型损失函数不变的基础上在每一层隐变量中都加入一个带有超参数的 TC 项作为正则化项，用来约束此层的隐变量，模型有几层就加几个。这样既可以使模型具有解耦表示学习能力，又不影响模型的重构质量。模型的损失函数包括：

模型的重构误差 L_{re}，为

$$L_{\mathrm{re}}=E_{q(\tilde{\boldsymbol{z}}_1|\boldsymbol{x})}[\ln p(\boldsymbol{x}\mid\tilde{\boldsymbol{z}}_1)] \tag{7}$$

KL 项 L_{KL}，为

$$L_{\mathrm{KL}}=\sum_{l=1}^{L}D_{\mathrm{KL}}\left(q(\boldsymbol{z}_l\mid\boldsymbol{x})\|p(\boldsymbol{z}_l)\right) \tag{8}$$

所有层的隐变量的 TC 项和控制其权重的超参数积的和 L_{regular}，为

$$L_{\mathrm{regular}}=\sum_{l=1}^{L}\left(\beta_l D_{\mathrm{KL}}\left(q(\boldsymbol{z}_l)\|\prod_j q(\boldsymbol{z}_{l_j})\right)\right) \tag{9}$$

其中：j 是每层隐变量的维度，$\beta_1,\beta_2,\cdots,\beta_L$ 是超参数。

TC-VLAE 模型损失函数为

$$L=L_{\mathrm{re}}-L_{\mathrm{KL}}-L_{\mathrm{regular}} \tag{10}$$

在本文模型训练时引入一种渐进式的训练方式，渐进式的训练方式将整个模型训练分为 L 步，即模型结构有几层隐变量训练就分为几步，每一步都是一个完整的模型训练过程，训练结束后得到模型参数。渐进式的训练方式如图 3 所示。

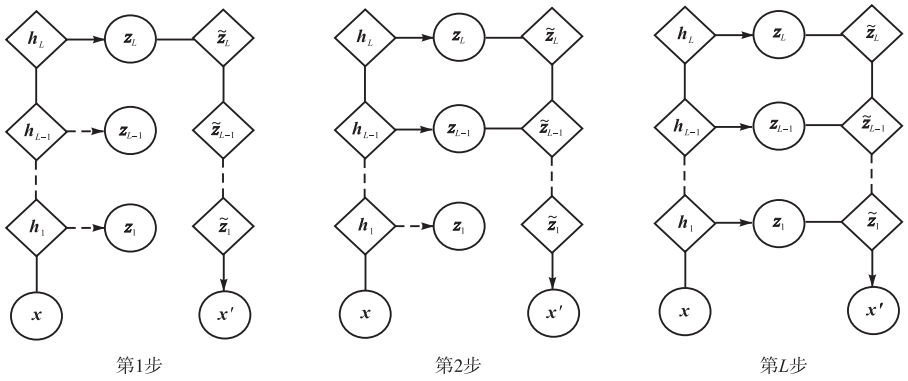


图 3　渐进式的训练方式

Fig. 3　Progressive training style

第 1 步，训练时随机初始化参数，在训练过程中只使用第 L 层的隐变量训练模型，第 1 层到第 L–1 层的隐变量不参加模型训练，训练完成后得到模型参数，缺少的参数使用随机数或 0 代替。第 2 步，训练时将第 1 步的结果作为本步的初始参数开始训练，训练时使用第 L 层与第 L–1 层的隐变量训练模型，第 1 层到第 L–2 层的隐变量不参加模型训练，训练完成后得到本步的模型参数，缺少的参数使用随机数或 0 代替。第 3 步，训练时将第 2 步的结果作为本步的初始参数开始训练，训练时使用第 L 层、第 L–1 层和第 L–2 层的隐变量，其他层的隐变量不参加训练。以此类推，直到最后一步所有的隐变量都参加模型训练并完成模型训练。

关于模型优化求解，首先对含有期望的损失函数使用蒙特卡洛采样和重参数化策略估计，然后再使用随机梯度下降的方法进行参数更新。

3　实　验

针对常见的解耦数据集设计对比实验，证明 TC-VLAE 模型与 TC-VAE 模型、VLAE 模型相比具有更强的解耦表示学习能力。具体包括 3 个实验：在 3Dshapes 数据集上比较 3 个模型 TC-VLAE、TC-VAE 和 VLAE 的解耦表示学习能力，在 3Dchairs 数据集和 CelebA 人脸数据集上验证 TC-VLAE 模型与单独基于非结构化先验模型在解耦表示学习方面的优势，在 dSprites 数据集上验证 TC-VLAE 模型与单独基于结构化先验模型在解耦表示学习方面的优势。最后，通过互信息差（mutual information gap，MIG）定量衡量模型在数据集上的解耦表示学习能力。

3.1　数据集

实验使用当下常见的用于评估模型解耦表示学习能力的数据集 3Dshapes 数据集[17]、3Dchairs 数据集[18]、CelebA 人脸数据集[19]和 dSprites 数据集[9]。

3Dshapes 数据集[18]是由 6 个真实独立的潜在因素生成的三维形状数据集，该数据集中的潜在因素有地板颜色、墙壁颜色、物体颜色、物体尺寸、物体形状和物体角度。该数据集由 480 000 张大小为（64，64，3）的 RGB 图像组成。数据集可视化如图 4 所示。

3Dchairs 数据集[19]由 1 000 个不同的 3D 椅子模型的渲染图像组成，是解耦表示学习研究中经常用到的数据集。数据集可视化如图 5 所示。

CelebA 人脸数据集[20]是香港中文大学开源的一个数据集，它包含了 10 177 个名人身份的 202 599 张大小为（64，64，3）的 RGB 图像。数据集可视化如图 6 所示。

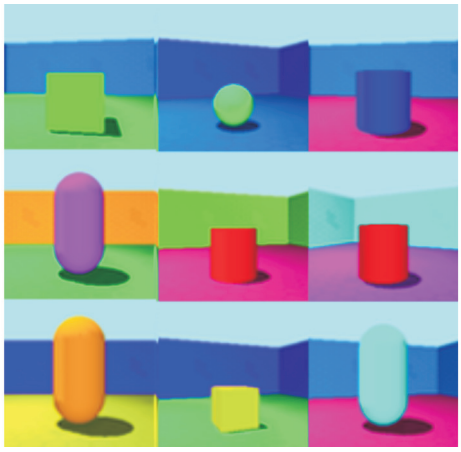


图 4　3Dshapes 数据集可视化
Fig. 4　Visualization of 3Dshapes datasets

图 5　3Dchairs 数据集可视化
Fig. 5　Visualization of 3Dchairs datasets

图 6　CelebA 数据集可视化
Fig. 6　Visualization of CelebA datasets

dSprites 数据集[9]是一个二维形状数据集，由 5 个真实独立的潜在因素生成。这些因素包括精灵的形状、比例、旋转和物体的横向、纵向的位置。数据集共有 737 280 张大小为(64, 64)的图像。数据集可视化如图 7 所示。

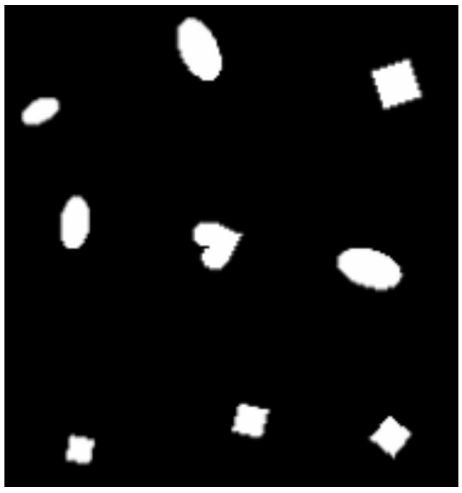


图 7　dSprites 数据集可视化
Fig. 7　Visualization of dSprites datasets

3.2　在 3Dshapes 数据集上的比较

TC-VLAE 模型在 3Dshapes 数据集的实验设置为：推理模型和生成模型均采用卷积神经网络，训练迭代次数为 15，批大小为 100，模型的神经网络架构为 3 层，每层隐空间维度为 3，训练中的超参数 $\beta=(8,8,8)$，模型训练过程中使用 Adam 优化训练，学习率为 0.000 1。实验结果如图 8 所示。

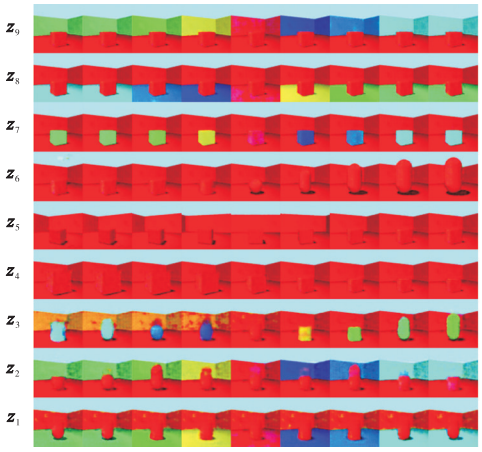


图 8　TC-VLAE 模型在 3Dshapes 数据集上解耦生成效果
Fig. 8　Disentanglement generation effect of TC-VLAE model on the 3Dshapes datasets

TC-VAE 模型[20]在 3Dshapes 数据集的实验设置为：推理模型和生成模型均采用卷积神经网络，训练迭代次数为 15，批大小为 100，隐空间维度为 9，训练中的超参数 $\beta=8$，模型训练过程中使用 Adam 优化训练，学习率为 0.000 1。实验结果如图 9 所示。

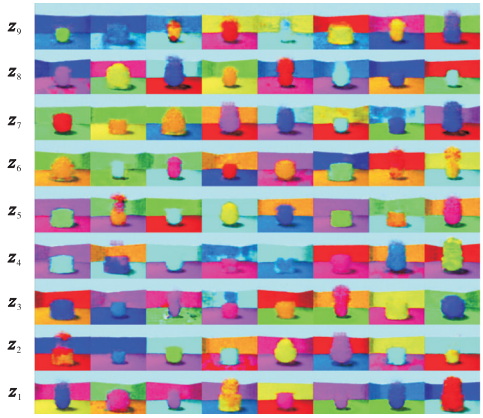


图 9　TC-VAE 模型在 3Dshapes 数据集上解耦生成效果

Fig. 9　Disentanglement generation effect of TC-VAE model on the 3Dshapes datasets

VLAE 模型[21]在 3Dshapes 数据集的实验设置为：推理模型和生成模型均采用卷积神经网络，训练迭代次数为 15，批大小为 100，模型的神经网络架构为 3 层，每层隐空间维度为 3，训练中的超参数 $\beta=(8,8,8)$，模型训练过程中使用 Adam 优化训练，学习率为 0.000 1。实验结果如图 10 所示。

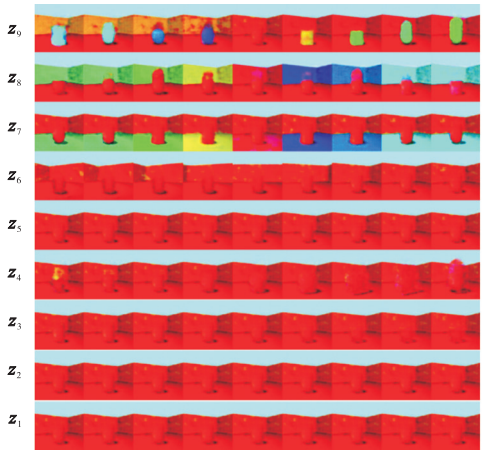


图 10　VLAE 模型在 3Dshapes 数据集上解耦生成效果

Fig. 10　Disentanglement generation effect of VLAE model on the 3Dshapes datasets

如图 8 所示，TC-VLAE 模型解耦生成效果自下而上分别为：z_1—z_3 对应第 1 层隐变量，这一层模型重构生成的图像并没有完全解耦，每一维度至少都有 2 个数据特征在变化，比如第 1 维度地板的颜色和物体的形状同时在变；z_4—z_6 对应第 2 层隐变量，这一层模型学习到比较低级层次的数据特征，实现了粗粒度解耦，比如第 4 维度学习到物体的尺寸逐渐由大变小，而其他的特征则相对不变，第 5 维和第 6 维分别学到物体朝向的角度和物体形状；z_7—z_9 对应第 3 层隐变量，这一层模型学习到高层次的数据特征，实现了细粒度的解耦，第 3 层的 3 个维度分别学习到物体的颜色、地板的颜色和背景的颜色。图 9 中 TC-VAE 模型在 3Dshapes 数据集上没有学习到有用的特征信息。图 10 中 VLAE 模型在 z_7—z_9 学习到有用的特征信息。

上述实验结果表明，基于结构化先验的 TC-VLAE 模型和 VLAE 模型在 3Dshapes 数据集上具有一定的解耦表示学习能力。这代表多层次的网络结构在形状规则的数据集（比如 3Dshapes 数据集）上有较强的解耦表示学习能力。TC-VLAE 模型在多层次的网络结构的基础上引入了基于非结构化的 TC 项，所以比 VLAE 模型能够学习到更多有用的特征信息，具有更强的解耦表示学习能力。

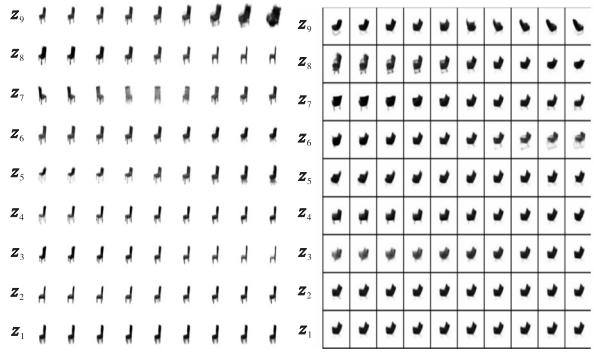
3.3　TC-VLAE 模型与基于非结构化先验模型 TC-VAE 的比较

3.3.1　在 3Dchairs 数据集上的比较

TC-VLAE 模型在 3Dchairs 数据集的实验设置为：推理模型和生成模型均采用卷积神经网络，训练迭代次数为 15，批大小为 100，模型的神经网络架构为 3 层，每层隐空间维度为 3，训练中的超参数 $\beta=(8,8,8)$，模型训练过程中使用 Adam 优化训练，学习率为 0.000 1。

TC-VAE 模型在 3Dchairs 数据集的实验设置为：推理模型和生成模型均采用卷积神经网络，训练迭代次数为 15，批大小为 100，模型隐空间维度为 9，训练中的超参数 $\beta=8$，模型训练过程中使用 Adam 优化训练，学习率为 0.000 1。

实验结果如图 11 所示。



（a）TC-VLAE 模型　　（b）TC-VAE 模型

图 11　在 3Dchairs 数据集上解耦生成效果

Fig. 11　Disentanglement generation effect on the 3Dchairs datasets

在图 11(a)中，z_1—z_3 对应第 1 层隐变量，这一

层解耦效果并不明显，更多的是没有发生变化；z_4 — z_6 对应第 2 层隐变量，这一层模型学习到一些数据特征，比如第 4 维学习到椅子的类型，第 5 维学习到椅子脚的类型，虽然第 6 维依旧学习到的是椅子的类型，但是该维度椅子类型的变化与第 4 维学习到的椅子类型并不相同；z_7 — z_9 对应第 3 层隐变量，这一层学习到更加复杂的数据特征，第 7 维和第 8 维都学习到椅子朝向的角度，但两者又有所区别，并不完全相同，第 9 维学习到椅子的尺寸。

在图 11(b)中 TC-VAE 模型的实验结果有 9 个维度，但并不是每个维度都学习到有意义的数据特征变化。虽然实验结果表示该模型在 3Dchairs 数据上能够学习到椅子朝向的方向(第 9 维)、椅子的类型(第 8 维)和椅子脚的类型(第 6 维)等多个数据特征，但是通过两图的对比可以看出，在 3Dchairs 数据集上，TC-VLAE 模型不仅比 TC-VAE 模型生成图片的质量高，而且能够学习到更加丰富且有意义的数据特征，具有更好的解耦表示学习能力。

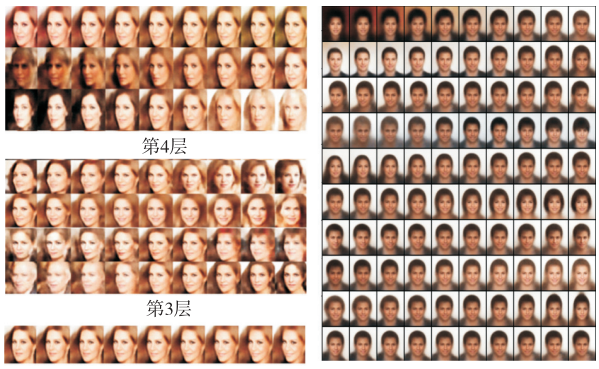
3.3.3　在 CelebA 人脸数据集上的比较

TC-VLAE 模型在 CelebA 人脸数据集的实验设置为：推理模型和生成模型均采用卷积神经网络，训练迭代次数为 10，批大小为 128，模型的神经网络架构为 4 层，每层隐空间维度为 7，训练中的超参数 $\beta=(5,5,5,5)$，模型训练过程中使用 Adam 优化训练，学习率为 0.000 1。

TC-VAE 模型在 CelebA 人脸数据集的实验设置为：推理模型和生成模型均采用卷积神经网络，训练迭代次数为 10，批大小为 128，模型的隐空间维度为 28，训练中的超参数 $\beta=5$，模型训练过程中使用 Adam 优化训练，学习率为 0.000 1。

实验结果如图 12 所示。本文模型在 CelebA 人脸数据集上的实验结果表明：模型在第 1 层没有学习到有意义的数据特征，在第 2 层学习到人微笑的特征变化，在第 3 层学习到人的性别(第 4 行，自上而下)、人的发型(第 3 行)以及人脸朝向的角度变化(第 1 行和第 2 行)，在第 4 层从上到下分别学习到头发的颜色、人脸的颜色和图片背景的颜色。对比本文模型与 TC-VAE 模型的实验结果可以发现，虽然 TC-VAE 模型也能够学习到人脸朝向的角度、背景颜色和人的性别等数据特征，但本文模型不仅生成的图像质量好，而且能够学习到更丰富、更细腻的数据特征。

上述实验结果表明，将结构化先验与非结构化先验相结合的 TC-VLAE 模型在 3Dchairs 数据集和 CelebA 数据集上的解耦效果比单独基于非结构化先验的 TC-VAE 模型更好。



(a) TC-VLAE 模型　　(b) TC-VAE 模型

图 12　在 CelebA 数据集上解耦生成效果

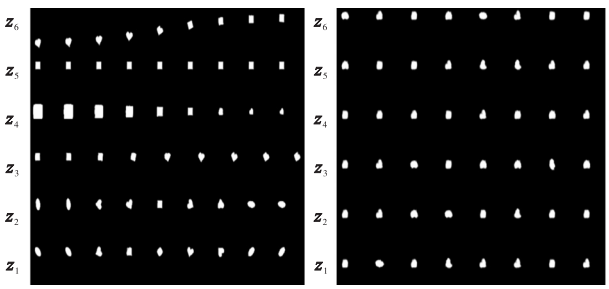
Fig. 12　Disentanglement generation effect on the CelebA datasets

3.4　TC-VLAE 模型与基于结构化先验模型 VLAE 的比较

TC-VLAE 模型在 dSprite 数据集的实验设置为：推理模型和生成模型均采用卷积神经网络，训练迭代次数为 10，批大小为 128，模型的神经网络架构为 3 层，每层隐空间维度为 2，训练中的超参数 $\beta=(5,5,5)$，模型训练过程中使用 Adam 优化训练，学习率为 0.000 1。

VLAE 模型在 dSprite 数据集的实验设置为：推理模型和生成模型均采用卷积神经网络，训练迭代次数为 10，批大小为 128，模型的神经网络架构为 3 层，每层隐空间维度为 2，训练中的超参数 $\beta=(5,5,5)$，模型训练过程中使用 Adam 优化训练，学习率为 0.000 1。

实验结果如图 13 所示。



(a) TC-VLAE 模型　　(b) VLAE 模型

图 13　dSprite 在数据集上解耦生成效果

Fig. 13　Disentanglement generation effect on the dSprite datasets

在图 13 中，自下而上分为 3 层，z_1 — z_2 对应第

1 层隐变量，z_3—z_4 对应第 2 层隐变量，z_5—z_6 对应第 3 层隐变量。结果表明：本文模型能够在第 1 层学习到物体的方向，在第 2 层学习到物体的类型和物体的大小，在第 3 层学习到物体的位置移动。图 13(b) 显示 VLAE 模型在此数据集上并没有明显的解耦表示学习能力。

上述实验结果表明，将结构化先验与非结构化先验相结合的 TC-VLAE 模型在 dSprite 数据集上的解耦效果要比单独基于结构化先验的 VLAE 模型更好。

3.5　定量的解耦评估指标

互信息差是一种可以定量衡量无监督模型的解耦表示学习能力的度量方式[11]，它是 0 ~ 1 之间的一个数，数值越大表示模型解耦表示学习能力越强。理论上一个无监督模型的解耦表示学习能力不会到 1。在解耦表示学习领域中，模型的解耦评估指标一直是一个重要的研究方向，许多研究者都提出了自己的解耦评估指标，比如 Kim 等[2]的工作，互信息差是其中使用较为广泛的一个。

将本文模型 TC-VLAE 和 TC-VAE 模型、VLAE 模型在 3Dshapes 数据集、3Dchairs 数据集、CelebA 人脸数据集和 dSprites 数据集上的互信息差进行汇总，结果见表 1。

表 1　各个模型在各个数据集上的互信息差

Tab. 1　Mutual information gap of each model on the datasets

模型	互信息差			
	3Dshapes	3Dchairs	CelebA	dSprite
TC-VAE	0.45	0.47	0.41	0.52
VLAE	0.39	0.26	0.19	0.21
TC-VLAE	0.79	0.65	0.59	0.68

TC-VLAE 模型比任何一个单思路的模型都具有更大的互信息差，说明 TC-VLAE 模型具有更强的解耦表示学习能力。

4　结　语

本文提出了 TC-VLAE 模型。TC-VLAE 模型是基于层次化的变分自编码模型，使模型实现不同层级间的特征解耦。在多层次模型结构中的每一层隐空间中都加入 TC 项作为正则化项，使模型实现各层级隐空间内部的特征解耦。训练时使用渐进式的训练方式充分发挥层次化网络模型的优势，同时使用蒙特卡洛法采样、重参数化策略以及随机梯度下降算法求解优化问题。本文在 4 个常用解耦数据集上进行对比实验，结果表明：将层次化结构与 TC 先验相结合的 TC-VLAE 模型的解耦表示学习能力比单一的层次化结构模型和只有 TC 先验的非结构化模型都更优异。

参考文献：

[1] KUMAR A，SATTIGERI P，BALAKRISHNAN A. Variational inference of disentangled latent concepts from unlabeled observations[EB/OL]. [2023-01-10]. https://arxiv.org/abs/1711.00848.

[2] KIM H，MNIH A. Disentangling by factorising[C]// PMLR. International Conference on Machine Learning. New York：PMLR，2018：2649-2658.

[3] BENGIO Y，COURVILLE A，VINCENT P，et al. Representation learning：a review and new perspectives[J]. IEEE Transactions on pattern analysis & machine intelligence，2013，35(8)：1798-1828.

[4] 文载道，王佳蕊，王小旭，等. 解耦表征学习综述[J]. 自动化学报，2022，48(2)：351-374.

[5] GOODFELLOW I，POUGET-ABADIE J，MIRZA M，et al. Generative adversarial nets[C]//MIT. Neural Information Processing Systems. Cambridge：MIT Press，2014.

[6] KINGMA D P，WELLING M. Auto-encoding variational bayes[EB/OL]. [2023-01-10]. https://arxiv.org/pdf/1312.6114v1.pdf.

[7] CHEN X，DUAN Y，HOUTHOOFT R，et al. InfoGAN：interpretable representation learning by information maximizing generative adversarial nets[EB/OL]. [2023-01-10]. https://doi.org/10.48550/arXiv.1606.03657.

[8] LEE W，KIM D，HONG S，et al. High-fidelity synthesis with disentangled representation[C]//ECCV. Computer Vision-ECCV 2020：16th European Conference. Berlin：Springer International Publishing，2020：157-174.

[9] HIGGINS I，MATTHEY L，PAL A，et al. β-VAE：learning basic visual concepts with a constrained variational framework[C]//ICLR. International conference on learning representations. Toulon：ICLR，2017.

[10] BURGESS C P，HIGGINS I，PAL A，et al. Understanding disentangling in beta-VAE[EB/OL]. [2023-01-10]. https://doi.org/10.48550/arXiv.1804.03599.

[11] CHEN R T Q，LI X，GROSSE R B，et al. Isolating sources of disentanglement in variational autoencoders [EB/OL]. [2023-01-10]. https://doi.org/10.48550/arXiv.1802.04942.

[12] MONTERO M L，LUDWIG C J H，COSTA R P，et al.

The role of disentanglement in generalisation[EB/OL]. [2023-01-10]. https://openreview.net/pdf?id = qbH974jKUVy.

[13] SØNDERBY C K, RAIKO T, MAALØE L, et al. Ladder variational autoencoders[EB/OL]. [2023-01-10]. https://arxiv.org/pdf/1602.02282.pdf.

[14] ZHAO S, SONG J, ERMON S. Learning hierarchical features from deep generative models[C]//ACM. Proceedings of the 34th International Conference on Machine Learning. Waco: ACM, 2017: 4091-4099.

[15] KARRAS T, AILA T, LAINE S, et al. Progressive growing of gans for improved quality, stability, and variation[EB/OL]. [2023-01-10]. http://arxiv.org/pdf/1710.10196.

[16] WANG Y, PERAZZI F, MCWILLIAMS B, et al. A fully progressive approach to single-image super-resolution [C]//IEEE. Proceedings of the IEEE conference on computer vision and pattern recognition workshops. New York: IEEE, 2018: 864-873.

[17] BURGESS C, KIM H. 3Dshapes dataset[EB/OL]. [2023-01-10]. https://github.com/deepmind/3dshapes-dataset/2018.

[18] AUBRY M, MATURANA D, EFROS A A, et al. Seeing 3D chairs: exemplar part-based 2D-3D alignment using a large dataset of cad models[C]//IEEE. Proceedings of the IEEE conference on computer vision and pattern recognition. New York: IEEE, 2014: 3762-3769.

[19] LIU Z, LUO P, WANG X, et al. Deep learning face attributes in the wild[C]//IEEE. Proceedings of the IEEE international conference on computer vision. New York: IEEE, 2015: 3730-3738.

[20] CHEN R T Q, LI X C, GROSSE R, et al. Isolating sources of disentanglement in variational autoencoders [EB/OL]. [2023-01-10]. https://doi.org/10.48550/arXiv.1802.04942.

[21] ZHAO S J, SONG J M, ERMON S. Learning hierarchical features from generative models[EB/OL]. [2023-01-10]. https://doi.org/10.48550/arXiv.1702.08396.

责任编辑：郎婧

（上接第 40 页）

dustry applications, 1987(5): 887-893.

[17] CRISTIANO R, PAGANO D J, HENAO M M. Multiple boundaries sliding mode control applied to capacitor voltage-balancing systems[J]. Communications in nonlinear science and numerical simulation, 2020, 91: 105430.

[18] 贾万水. NPC 三电平逆变器中点电压平衡研究[D]. 株洲：湖南工业大学, 2022.

[19] 夏玉政, 胡海兵, 邢岩. 谐波补偿下考虑中点平衡的等效三电平调制[J]. 电力电子技术, 2021, 55(4): 133-136.

[20] HASSAN M S, ABDELHAKIM A, SHOYAMA M, et al. On-the-analysis and reduction of common-mode voltage of a single-stage inverter through control of a four-leg-based topology[J]. International journal of electrical power & energy systems, 2021, 127: 106710.

[21] YUE Y F, XU Q M, GUO P, et al. Capacitor voltage predictor-corrector balancing approach with single sensor for single-phase modular multilevel converter[J]. International journal of electrical power & energy systems, 2021, 129: 106729.

[22] OZDEMIR S, ALTIN N, SEFA I, et al. Super twisting sliding mode control of three-phase grid-tied neutral point clamped inverters[J]. ISA Transactions, 2022, 125: 547-559.

[23] USHA S, GEETHA A, THENTRAL T M T, et al. Mitigation of common mode voltage in five phase multilevel inverter[J]. Materials today: proceedings, 2021, 45: 1761-1769.

[24] 胡昭, 潘三博. 一种抑制共模电压的 T 型三电平逆变器调制策略[J]. 上海电机学院学报, 2022, 25(3): 125-131.

[25] HAKAMI S S, LEE K B. Enhanced predictive torque control for three-level NPC inverter-fed PMSM drives based on optimal voltage magnitude control method[J]. IEEE Transactions on power electronics, 2022, 38(3): 3725-3738.

责任编辑：周建军