中图法分类号:TP37　　**文献标识码**: A　　**文章编号**: 1006-8961(2023)04-0903-32

论文引用格式:Li Y T, Xiao J, Liao L, Wang Z, Chen W Y and Wang M. 2023. A review of disentangled representation learning for visual data processing and analysis. Journal of Image and Graphics, 28(04):0903-0934(李雅婷, 肖晶, 廖良, 王正, 陈文益, 王密. 2023. 面向视觉数据处理与分析的解耦表示学习综述. 中国图象图形学报, 28(04):0903-0934)[DOI:10.11834/jig.211261]

面向视觉数据处理与分析的解耦表示学习综述

李雅婷[1],肖晶[1*],廖良[2],王正[1],陈文益[1],王密[3]

1. 武汉大学计算机学院国家多媒体软件工程技术研究中心,武汉　430072; 2. 日本国立信息学研究所数字内容和媒体科学研究部,东京　101-8430,日本; 3. 武汉大学测绘遥感信息工程国家重点实验室,武汉　430079

摘　要: 表示学习是机器学习研究的核心问题之一。机器学习算法的输入表征从过去主流的手工特征过渡到现在面向多媒体数据的潜在表示,使算法性能获得了巨大提升。然而,视觉数据的表示通常是高度耦合的,即输入数据的所有信息成分被编码进同一个特征空间,从而互相影响且难以区分,使得表示的可解释性不高。解耦表示学习旨在学习一种低维的可解释性抽象表示,可以识别并分离出隐藏在高维观测数据中的不同潜在变化因素。通过解耦表示学习,可以捕获到单个变化因素信息并通过相对应的潜在子空间进行控制,因此解耦表示更具有可解释性。解耦表征可用于提高样本效率和对无关干扰因素的容忍度,为数据中的复杂变化提供一种鲁棒性表示,提取的语义信息对识别分类、域适应等人工智能下游任务具有重要意义。本文首先介绍并分析解耦表示的研究现状及其因果机制,总结解耦表示的3个重要性质。然后,将解耦表示学习算法分为4类,并从数学描述、类型特点及适用范围3个方面进行归纳及对比。随后,分类总结了现有解耦表示工作中常用的损失函数、数据集及客观评估指标。最后,总结了解耦表示学习在实际问题中的各类应用,并对其未来发展进行了探讨。

关键词: 解耦表示学习;视觉数据;潜在表征;变化因素;潜空间

A review of disentangled representation learning for visual data processing and analysis

Li Yating[1], Xiao Jing[1*], Liao Liang[2], Wang Zheng[1], Chen Wenyi[1], Wang Mi[3]

1. National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University, Wuhan 430072, China; 2. Digital Content and Media Sciences Research Division, National Institute of Informatics, Tokyo 101-8430, Japan; 3. State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China

Abstract: Representation learning is essential for machine learning technique nowadays. The transition of input representations have been developing intensively in algorithm performance benefited from the growth of hand-crafted features to the representation for multi-media data. However, the representations of visual data are often highly entangled. The interpretation challenges are to be faced because all information components are encoded into the same feature space. Disentangled representation learning (DRL) aims to learn a low-dimensional interpretable abstract representation that can sort the multiple factors of variation out in high-dimensional observations. In the disentangled representation, we can capture and manipulate the information of a single factor of variation through the corresponding latent subspace, which makes it more

收稿日期:2022-01-21;**修回日期**:2022-05-16;**预印本日期**:2022-05-23

*通信作者:肖晶　jing@whu.edu.cn

基金项目:湖北省自然科学基金项目(2020CFA001);湖北省重点研发计划项目(2020BIB006)

Supported by:Natural Science Foundation of Hubei Province, China (2020CFA001); Key R&D Program of Hubei Province, China (2020BIB006)

interpretable. DRL can improve sample efficiency and tolerance to the nuisance variables and offer robust representation of complex variations. Their semantic information is extracted and beneficial for artificial intelligence (AI) downstream tasks like recognition, classification and domain adaptation. Our summary is focused on brief introduction to the definition, research development and applications of DRL. Some of independent component analysis (ICA)-nonlinear DRL researches are covered as well since the DRL is similar to the identifiability issue of nonlinear independent component analysis (nonlinear ICA). The cause and effects mechanism of DRL as high-dimensional ground truth data is generated by a set of unobserved changing factors (generating factors). The DRL can be used to model the factors of variation in terms of latent representation, and the observed data generation process is restored. We summarize the key elements that a well-defined disentangled representation should be qualified into three aspects, which are 1) modularity, 2) compactness, and 3) explicitness. First, explicitness consists of the two sub-requirements of completeness and informativeness. Then, current DRL types are categorized into 1) dimension-wise disentanglement, 2) semantic-based disentanglement, 3) hierarchical disentanglement, and 4) nonlinear ICA four types in terms of its formulation, characteristics, and scope of application. Dimension-wise disentanglement is assumed that the generative factors are solely and each dimension of latent vector can be separated and mapped, which is suitable for learning the disentangled representation of simple synthetic visual data. Semantic-based disentanglement is hypnotized that some semantic information is solely as well. The generative factors are group-disentangled in terms of specific semantics and they are mapped to different latent spaces, which is suitable for complicated ground truth data. Hierarchical disentanglement is based on the assumption that there is a correlation between generative factors at different levels of abstraction. The generative factors are disentangled by group from the bottom up and they can be mapped to latent space of different semantic abstraction levels to form a hierarchical disentangled representation. Nonlinear ICA provides an identifiable method for observed data-mixed disentangling unknown generative factors through a nonlinear reversible generator. For the motivation of loss functions, the loss functions can be commonly used in disentangled representation learning, which are grouped into three categories: 1) modularity constraint: a single latent variable-constrained in the disentangled representation to capture only a single or a single group of factors of variation, and it promotes the separation of factors of variation mutually; 2) explicitness constraint: current latent variable of the latent representation is activated to encode the ground truth of the corresponding generating factor effectively, and the entire latent representation contains complete information about all generative factors; and 3) multi-purpose constraint: loss-related can optimize multiple disentangled representation, including modularity, compactness, and explicitness of the disentangled representation at the same time. The model-relevant can combine multiple loss constraint terms to form the final hybrid objective function. We compare the scope of application and limitations of each type of loss functions and summarize the classical disentangled representation works using the hybrid objective function further.

Key words: disentangled representation learning; visual data; latent representation; factors of variation; latent space

0 引 言

表示学习(representation learning)是机器学习和计算机视觉的基本研究问题之一(Bengio 等,2013)。现实世界的感官数据,如图像、视频和音频,往往以高维的形式存在。表示学习将这些数据映射到一个低维潜空间(latent space)中,从高维数据中学习可解释的潜在表示(latent representation),使其更容易为下游任务(如分类和检测)提供有语义的信息。然而,深度学习方式下学到的表示总是高度耦合的,即观测数据的各类因素信息被混合编码到同一个空间而难以区分,使得学到的表征很难被阐释。解耦表示学习(disentangled representation learning,DRL)正是为了解决这一问题,其目的是学习一种结构化表示,以识别并分离出隐藏在观测数据中的潜在可解释因素(Bengio 等,2013)。

对于解耦表示(disentangled representation)的概念,目前还没有一个被广泛接受的正式定义。 Higgins 等人(2018)使用物理学的对称群表示理论(symmetry group representation learning),将解耦表示定义为一个可被分解为若干个子空间的向量,其中每个子空间都与一个唯一的对称变换(symmetry transformation)相匹配,并可独立地进行变换。Ben-

gio 等人(2013)将解耦表示定义为表征的单个潜在单元只对一个生成因素的变化很敏感而对其他因素的变化则保持相对不变。虽然没有一个广泛采用的正式定义,但在直观认知上达成一致的是希望学习到一种关于观测数据的可解释的结构化低维潜在表示,其中各潜变量分别恢复出不同的真实生成因素。解耦表示学习的目的和非线性独立成分分析(nonlinear independent component analysis,nonlinear ICA)的可识别性问题(Khemakhem 等,2020)十分接近,因此本文在第2节中也涵盖了一部分基于非线性独立成分分析的解耦工作。

解耦表示学习的早期工作(Kulkarni 等,2015;Reed 等,2014;Yang 等,2015;Zhu 等,2014)大多采用有监督学习方法,需要大量生成因素的标签信息作为监督信号,指导模型学习数据中潜在的变化因素。然而标签制作需要耗费大量人力,因此在解耦表示学习发展的中后期,提出了许多无监督解耦表示模型。即解耦表示模型不需要任何生成因素的标签信息作为标签,仅通过损失函数的约束即可实现生成因素间的分离。例如,β-VAE(variational autoencoders)(Higgins 等,2017)通过约束潜在表示的信息瓶颈的容量以分离不同的变化因素;Zwicker 等人(2018)使用交叉重建损失实现特征解耦;IIAE(interaction information autoencoder)(Hwang 等,2020)通过互信息约束分离潜在表示。然而,现实世界中数据的生成因素丰富且交互复杂,缺乏显式性标签做监督、只靠损失函数约束以实现解耦的无监督学习算法并不能可控地分离出生成因素。因此,近年来,仅需少量显性或者隐性生成因素知识的弱监督和自监督学习成为解耦表示学习的主流方式。例如,FineGAN(fine-grained generative adversarial network)(Singh 等,2019)仅需极少量显性生成因素知识,即一个边界框(bounding box)作为监督信号,便可同时解耦出真实图像中背景、对象形状、姿势和外貌纹理等4种不同的变化因素。

解耦表示应用于下游任务,通过忽略观测数据中与任务无关的生成因素信息和随机噪声,可增强算法的鲁棒性和对噪声的容忍性;通过保留并分离出与任务相关的不同变化因素,可以降低样本复杂度、提供可解释性和精确可控性、提高识别分类的准确性,对识别分类(Bai 等,2020a;Liu 等,2020;Zhao 等,2018)、域适应(Baktashmotlagh 等,2018;Cai 等,2019;Peng 等,2019)、图像生成与处理(Gilbert 等,2018;Liu 等,2018b;Zhu 等,2018)、公平机器学习(Creager 等,2019;Locatello 等,2019a)和推理(Van Steenkiste 等,2019)等深度学习任务具有重要意义。

1 解耦表示的因果机制及特性

1.1 因果机制

如图1所示,在解耦表示中,通常假设真实世界的高维观测数据$\boldsymbol{x}$是由一组未被观测到的变化因素(生成因素)$\boldsymbol{G}=\{g_1,g_2,\cdots,g_m\}$产生的。解耦表示学习通过寻找一个潜在的表示$\boldsymbol{z}$来建模变化因素,并还原观测数据的生成过程。通常假设观测数据$\boldsymbol{x}$是由一个两步生成过程产生的。首先,从分布$P(\boldsymbol{z})$中采样一个多元潜在随机向量$\boldsymbol{z}=[z_1,z_2,\cdots,z_n]$。因为随机向量$\boldsymbol{z}$的一个取值对应于潜在表示空间的单个数据点,所以随机向量$\boldsymbol{z}$也称为潜码(latent code)。理想条件下,z_i对应于观测数据中在语义上具有物理意义的一个或一组变化因素。然后,观察值$\boldsymbol{x}$从条件分布$P(\boldsymbol{x}|\boldsymbol{z})$中采样得到。其关键思想是高维数据$\boldsymbol{x}$可以由维度低得多、语义丰富的潜向量$\boldsymbol{z}$解释,且两者之间存在一定的映射关系。解耦表示的目的就是地捕捉到$\boldsymbol{x}$中所有真实的潜在变化因素,从而在下游任务中更容易提取有用的信息。

1.2 解耦表征的特性

近年来,部分研究者就规范的解耦表示应该符合哪些标准进行了深入探讨。Eastwood 和 Williams(2018)、Ridgeway 和 Mozer(2018)以及 Higgins 等人(2018)都认为解耦表示需要满足模块化、紧凑性和明确性这3个理想属性。然而,上述3个工作均基于维度解耦(Kim 和 Mnih,2018)方法对解耦表示的3个标准要求进行定义和阐述,存在一定的局限性。为了解决这一问题,本文将维度解耦、语义解耦和非线性独立成分分析均考虑在内,对这3个属性进行一定的延伸扩充。

1)模块化(modularity)指潜在表示的潜变量或分量最多只能捕获一个生成因素。每个潜码或者其每个维度(通道)只与一个变化因素相关,不受其他生成因素变化的影响,隐性地要求各潜变量可分离。

2)紧凑性(compactness)指尽可能用维度更加紧凑的潜空间来完整地表示单个真实生成因素的信

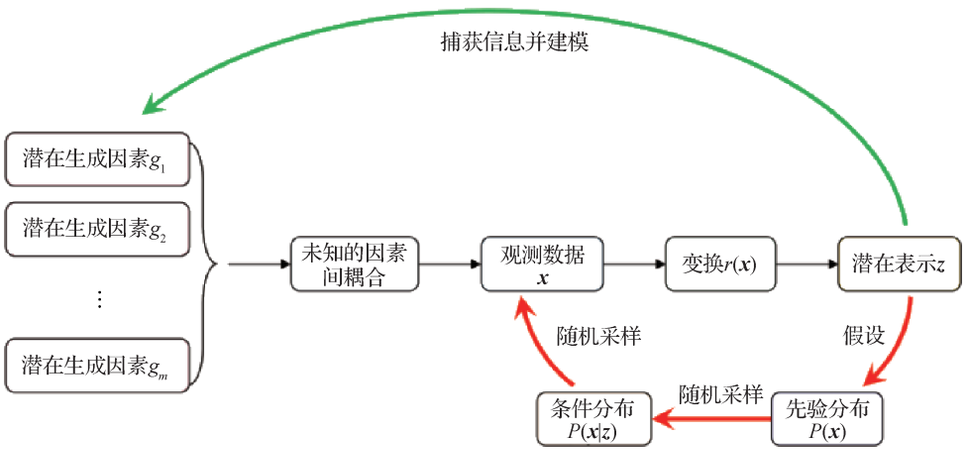


图1　解耦表示学习因果机制示意图

Fig. 1　Illustration of DRL causal mechanism

息。如果潜空间的维度过大，则噪声信息或冗余信息可能会被编码进潜在表示中。

3）明确性（explicitness）指潜变量模型可以从解耦表示中解码恢复出所有生成因素的信息。Higgins等人（2018）认为明确性是3个要求中最强的一个，因为明确性包含了完备性（completeness）和信息性（informativeness）两个子要求。完备性指解耦表示必须完整表达观测数据的所有变化因素信息。信息性也称成可解码性，指解耦特征通过模型可以恢复出生成因素的值。

2　解耦表示学习方法

根据生成因素间的关系和潜在表示中各潜变量的关系将常见的解耦表示工作总结为维度解耦、语义解耦、层级解耦和非线性独立成分分析四类，如图2所示。对于每类解耦表示方式，首先从数学角度描述其概率模型，然后探讨其特点和适用范围，并根据变分自动编码器（variational auto-encoder，VAE）（Kingma和Welling，2022）、生成对抗网络（generative adversarial network，GAN）（Goodfellow等，2014）和流模型（flow-based model）（Rezende等，2015）等生成模型，对该类解耦表示中的典型工作进行介绍。

2.1　维度解耦

2.1.1　问题描述

假设观测数据生成因素$\boldsymbol{G}=\{g_1,g_2,\cdots,g_m\}$互相独立，维度解耦（dimension-wise disentanglement）表示学习用一个n维的潜向量$\boldsymbol{z}$作为高维观测数据$\boldsymbol{x}\in\boldsymbol{X}$的潜在表示，捕获生成因素的信息。观测数据和潜向量的联合分布$p_\theta(\boldsymbol{x},\boldsymbol{z})$可表达为

$$\begin{aligned}p_\theta(\boldsymbol{x},\boldsymbol{z})&=p_\theta(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})\\p(\boldsymbol{z})&=N(\boldsymbol{z}|0,\sigma^2\boldsymbol{I})\end{aligned}\tag{1}$$

式中，潜变量$\boldsymbol{z}$的先验分布$p(\boldsymbol{z})$常假设为各个方向方差都一样的多维高斯分布，σ为标准差，条件概率$p_\theta(\boldsymbol{x}|\boldsymbol{z})$用参数为$\theta$的生成模型建模。

以观测值$\boldsymbol{x}$为条件的潜变量$\boldsymbol{z}$的后验分布则可用推论模型表示为

$$p_\theta(\boldsymbol{z}|\boldsymbol{x})=\frac{p_\theta(\boldsymbol{x},\boldsymbol{z})}{\int p_\theta(\boldsymbol{x},\boldsymbol{z})\mathrm{d}\boldsymbol{z}}\tag{2}$$

在VAE（Kingma和Welling，2022）模型中，由于潜变量$\boldsymbol{z}$的真实后验分布$p_\theta(\boldsymbol{x}|\boldsymbol{z})$难以实现，所以引入变分推论模型$q_\phi(\boldsymbol{x}|\boldsymbol{z})$来近似真实的后验分布。推论模型在观测数据$\boldsymbol{x}$的边际似然性上的变分下界（evidence lower bound，ELBO）为

$$L_{\mathrm{ELBO}}=E_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z})\right]-KL\left(q_\phi(\boldsymbol{z}|\boldsymbol{x})|p(\boldsymbol{z})\right)\tag{3}$$

式中，ϕ是推论模型参数，$KL(\cdot)$是KL（Kullback-Leibler）散度函数。

2.1.2　特点及适用性

如图2（a）所示，维度解耦表示学习通常仅用一个低维的潜向量$\boldsymbol{z}=[z_1,z_2,\cdots,z_n]$（也称为潜码）来表达高维观测数据的各种潜在变化因素g。理想情况下，潜向量$\boldsymbol{z}$的各个维度互相独立，每个维度z_i是一个单独的潜变量，代表观测数据的一种变化因素$\boldsymbol{g}_j$。维度解耦表示学习尝试完全分离所有的生成因素并映射到潜空间的各个维度上，因此这种类型是一种细粒度的解耦表示学习。

现有的维度解耦工作主要是基于VAE的方法

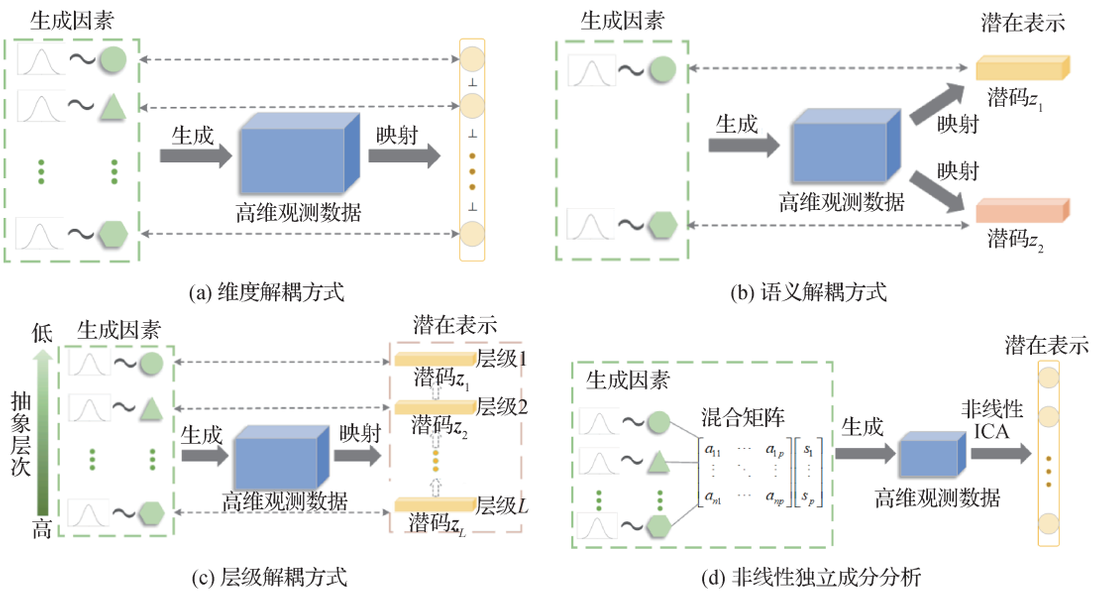


图 2　4 种类型的解耦表示学习模式示意图

Fig. 2　Paradigm diagram of four types of DRL((a)dimension-wise disentanglement; (b)semantics-based disentanglement; (c)hierarchical disentanglement; (d)nonlinear ICA)

(Achille 等，2018; Esmaeili 等，2019; Estermann 等，2020; Ma 等，2019; Painter 等，2020)，基于 GAN 的方法(Chen 等，2016; Zwicker 等，2018; Nie 等，2020b; Peebles 等，2020; Zhu 等，2020a)。维度解耦表示模型(Burgess 等，2018; Chen 等，2019; Dupont，2018; Higgins 等，2017; Kumar 等，2018)之间的差异在于目标函数，即实现传统 ELBO 不同形式的变体。例如，β-VAE(Higgins 等，2017)通过一个大于 1 的系数 β 惩罚 KL 散度项，约束潜变量 z 的信息瓶颈，鼓励潜向量维度间解耦。AnnealedVAE(Burgess 等，2018)在 β-VAE(Higgins 等，2017)的基础上，在 KL 散度项内引入变量 C 作为目标，进一步约束潜信息编码容量。β-TCVAE (total correlation variational autoencoder)(Chen 等，2019)将 KL 散度项分解为互信息、总相关(total correlation，TC)和逐维度 KL 散度三项，并分别用 α,β,γ 系数进行惩罚。基于 GAN(Goodfellow 等，2014)的代表工作是 InfoGAN(information-theoretic extension to the generative adversarial network)(Chen 等，2016)，该算法通过最大化观测数据和潜码之间的互信息学习鼓励解耦表达。

维度独立的解耦表示学习适用于简单的合成数据集和解耦数据集。合成数据的潜在变化因素个数少，而且每个生成因素分布简单。因此，以多元各异性高斯分布建模分布的潜变量适用于拟合简单合成数据的生成因素，且一个潜向量的单个维度空间足以编码合成数据的一个生成因素的信息。

2.2　语义解耦

2.2.1　问题描述

假设观测数据生成因素 $\boldsymbol{G}=\{g_1,g_2,\cdots,g_m\}$ 互相独立是一种理想的简单情况。现实世界中，复杂数据产生于许多生成因素的丰富交互中(Bengio 等，2013)。因此，生成因素并不是互相独立而是成群独立。假设生成因素 $\boldsymbol{G}$ 可划分为两个子集 $\boldsymbol{G}^{(1)}=\{g_1,g_2,\cdots,g_m\}$ 和 $\boldsymbol{G}^{(2)}=\{g_1,g_2,\cdots,g_m\}$，子集内的生成因素存在依赖关系，但子集之间互相独立，即

$$\boldsymbol{G}=\boldsymbol{G}^{(1)}\cup\boldsymbol{G}^{(2)},\boldsymbol{G}^{(1)}\perp\boldsymbol{G}^{(2)} \tag{4}$$

式中，$\perp$ 表示特征正交潜在表示。

语义解耦(semantics-based disentanglement)表示学习用潜向量 z_1 和 z_2 作为高维观测数据 $\boldsymbol{x}$ 的潜在表示，分别捕获生成因素子集 $\boldsymbol{G}^{(1)}$ 和 $\boldsymbol{G}^{(2)}$ 的信息。$\boldsymbol{G}^{(1)}$ 和 $\boldsymbol{G}^{(2)}$ 通常在语义上具有不同的物理含义。假设 $\boldsymbol{G}^{(1)}$ 具有语义 s_1，$\boldsymbol{G}^{(2)}$ 具有语义 s_2，则潜向量 z_1 和 z_2 分别对应语义信息 s_1 和 s_2。此时，观测数据和潜向量的联合分布 $p_\theta(\boldsymbol{x},z_1,z_2)$ 可表达为

$$p_\theta(\boldsymbol{x},z_1,z_2)=p_\theta(\boldsymbol{x}\mid z_1,z_2)\,p(z_1)\,p(z_2) \tag{5}$$

式中，潜变量 z_1 和 z_2 的先验分布 $p(z_1)$ 和 $p(z_2)$ 常被假设为高斯分布或者混合高斯分布。$p_\theta(\boldsymbol{x}\mid z_1,z_2)$ 代表参数 θ 的生成模型，根据输入的潜变量 z_1 和 z_2 得到 $\boldsymbol{x}$ 的条件分布。

以观测值$\boldsymbol{x}$为条件的潜变量$\boldsymbol{z}_1$和$\boldsymbol{z}_2$的后验分布则可用推论模型表示为

$$p_\theta(\boldsymbol{z}_1|\boldsymbol{x})=\frac{p_\theta(\boldsymbol{x},\boldsymbol{z}_1)}{\int p_\theta(\boldsymbol{x},\boldsymbol{z}_1)\mathrm{d}\boldsymbol{z}_1}$$
$$p_\theta(\boldsymbol{z}_2|\boldsymbol{x})=\frac{p_\theta(\boldsymbol{x},\boldsymbol{z}_2)}{\int p_\theta(\boldsymbol{x},\boldsymbol{z}_2)\mathrm{d}\boldsymbol{z}_2} \tag{6}$$

在基于VAE方法的语义解耦表示学习中(Hsieh等,2018;Hsu等,2017;Zhu等,2020b),推论模型在观测数据$\boldsymbol{x}$的边际似然性上的变分下界为

$$L_{\mathrm{ELBO}}=E_{q_\phi(\boldsymbol{z}_1,\boldsymbol{z}_2|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x}|\boldsymbol{z}_1,\boldsymbol{z}_2)\right]-$$
$$KL\left(q_\phi(\boldsymbol{z}_1|\boldsymbol{x})|p(\boldsymbol{z}_1)\right)-KL\left(q_\phi(\boldsymbol{z}_2|\boldsymbol{x})|p(\boldsymbol{z}_2)\right) \tag{7}$$

2.2.2 特点及适用性

语义解耦方式常见于解耦学习的实际应用问题中,也是现在解耦表示学习的主流方式。如图2(b)所示,研究者根据实际问题和数据的特性,尝试将生成因素$\boldsymbol{G}$解耦成两个或多个互相独立且具有不同语义的子集$\boldsymbol{G}^{(1)},\boldsymbol{G}^{(2)},\cdots,\boldsymbol{G}^{(k)}$,并用两个或多个潜码$\boldsymbol{z}_1,\boldsymbol{z}_2,\cdots,\boldsymbol{z}_k$分别表达这些生成因素子集。理想情况下,各潜码之间互相独立,一个潜码$\boldsymbol{z}_i$是一个单独的潜变量,编码具有语义s_i的生成因素子集$\boldsymbol{G}^{(i)}$的信息。所以,此时潜码也可看做语义特征。语义独立的解耦表示学习并不完全解耦所有的生成因素,只挑选出对下游任务起关键性作用的几类语义信息,将所有生成因素按语义逐群分离并映射到不同的潜空间,因此属于一种粗粒度的解耦表示学习。

语义解耦表示学习通常按"内容—风格"或者"相关性—无关性"二元语义模式将生成因素逐群解耦。为了使每个潜码捕获相应语义的生成因素信息,模型通常需要辅助任务或(伪)标签等辅助信号监督解耦表示编码的语义信息。在实现上,生成模型类型丰富,有VAE、GAN、AE(autoencoder)(Hinton和Salakhutdinov,2006)和Siamese network(Bromley等,1993)。基于VAE的方法(Bepler等,2019;Bouchacourt等,2018;Cai等,2019;Massagué等,2020;Detlefsen和Hauberg,2019;Ding等,2020;Fraccaro等,2017;Grathwohl和Wilson,2016),保持ELBO目标函数不变,通过改进推论模型和生成模型促进表征解耦。基于GAN的方法(Bai等,2020b;Bi等,2019;Duan等,2020;Gowal等,2020;Jiang等,2019;Lai等,2020;Ruan等,2020;Sun等,2019b;Zhang等,2020)常在像素空间和潜空间均引入判别器,保证生成数据的真实性和潜变量的语义性(Lee等,2018;Tulyakov等,2018;Xiao等,2019a),有时采用噪声注入机制使生成数据具有多样性(Gonzalez-Garcia等,2018;Li等,2020d)。基于AE的方法(Hadad等,2018;Li等,2020e;Liu等,2018a;Lorenz等,2019;Ma等,2018;Niu等,2020;Tsai等,2019)使用自编码器从数据空间自动分离生成因素并在潜空间编码有效信息,然后再由潜空间映射回数据空间。基于孪生网络(siamese network)(Bromley等,1993)的方法(Nie等,2020a;Wu和Lu,2020;Yin等,2019)用两个共享权重的相同网络,分辨输入的数据对,鼓励解耦特征发现并编码数据对间的相似性语义。

语义独立的解耦表示学习适用于复杂的真实世界数据,并用以解决下游任务。真实世界数据的生成因素多、交互丰富、分布复杂。按语义划分的逐群解耦比完全解耦可行性更高。此外,单个潜变量的空间是整个潜码而非其中的一个维度,足以捕获单个生成因素子集的完整信息。最终学习到的解耦表示提炼了对下游任务的关键信息,丢弃了无关的干扰信息,提高样本的效率,增强鲁棒性。

2.3 层级解耦

2.3.1 问题描述

假设观测数据的生成因素$\boldsymbol{G}=\{g_1,g_2,\cdots,g_m\}$间存在连续依赖关系,且它们处在不同抽象水平。层级解耦(hierarchical disentanglement)表示学习用一个潜向量$\boldsymbol{z}$学习高维观测数据$\boldsymbol{x}\in\boldsymbol{X}$的潜在层级表示,并将向量$\boldsymbol{z}$分解为$L$个来自不同语义抽象水平的潜变量$\{\boldsymbol{z}_1,\boldsymbol{z}_2,\cdots,\boldsymbol{z}_L\}$。潜变量$\boldsymbol{z}_l(l=1,2,\cdots,L)$的层级数$l$越大,代表的语义抽象水平越高。相反,潜变量$\boldsymbol{z}_i$的层级数$l$越小,代表的语义抽象水平越低。层级解耦表示学习模型主要有堆叠层级和架构层级两类(Zhao等,2017)。

堆叠层级模型(Gulrajani等,2016;Hsu等,2017;Rezende等,2014;Sønderby等,2016)认为潜变量的先验分布之间存在层级依赖关系,通过堆叠一系列生成模型来实现层级条件依赖。观测数据和潜向量的联合分布$p(\boldsymbol{x},\boldsymbol{z})$用链式法则可表达为

$$p(\boldsymbol{x},\boldsymbol{z})=p(\boldsymbol{x}|\boldsymbol{z}_1,\boldsymbol{z}_2,\cdots,\boldsymbol{z}_l)\prod_{l=1}^{L-1}p(\boldsymbol{z}_l|\boldsymbol{z}_{>l}) \tag{8}$$

式中,$\boldsymbol{z}_{>l}$指$\boldsymbol{z}_{l+1},\cdots,\boldsymbol{z}_L$,$p(\boldsymbol{z}_l|\boldsymbol{z}_{>l})$表示每个潜变量$\boldsymbol{z}_l$直

接依赖于比其抽象水平高的所有潜变量$\boldsymbol{z}_{>l}$。对于概率分布$p(\boldsymbol{x}\mid\boldsymbol{z}_{>0})$和$p(\boldsymbol{z}_l\mid\boldsymbol{z}_{>l})$，$l=1,2,\cdots,L-1$，每个条件概率分布用一个生成模型建模。所以联合分布$p(\boldsymbol{x},\boldsymbol{z})$是由$L$个模型堆叠形成的层级解耦模型。

架构层级模型（Li等，2020g；Zhao等，2017）并不假设潜变量之间存在层级依赖关系，L个潜变量的先验分布互相独立，用一个生成模型的不同网络层建模不同抽象水平的潜变量的先验分布。每个潜变量$\boldsymbol{z}_l$代表某个抽象层次的生成因素，这些生成因素不被其他抽象层次所捕获。因此，架构层级模型的生成模型为

$$p(\boldsymbol{x},\boldsymbol{z})=p\left(\boldsymbol{x}\mid\boldsymbol{z}_1,\boldsymbol{z}_2,\cdots,\boldsymbol{z}_L\right)\prod_{l=1}^{L}p\left(\boldsymbol{z}_l\right)\tag{9}$$

堆叠层级模型和架构层级模型的区别如图3所示。圆圈代表随机结点，菱形代表确定性计算节点，带箭头实线代表条件概率，无箭头实线代表确定性映射，虚线代表匹配潜变量先验分布所需的正则化操作。然而，堆叠层级模型和架构层级模型均不假设L个潜变量的后验分布之间层级依赖，每个潜变量$\boldsymbol{z}_l$的后验分布只依赖于观测数据$\boldsymbol{x}$的某个抽象水平$s_l(x)$。因此，两类层级解耦表示的推论模型可表示为

$$\boldsymbol{h}_l=S_l(\boldsymbol{x})\tag{10}$$

$$q\left(\boldsymbol{z}_1,\boldsymbol{z}_2,\cdots,\boldsymbol{z}_L\mid\boldsymbol{x}\right)=\prod_{l=1}^{L}q\left(\boldsymbol{z}_l\mid s_l\left(\boldsymbol{x}\right)\right)\tag{11}$$

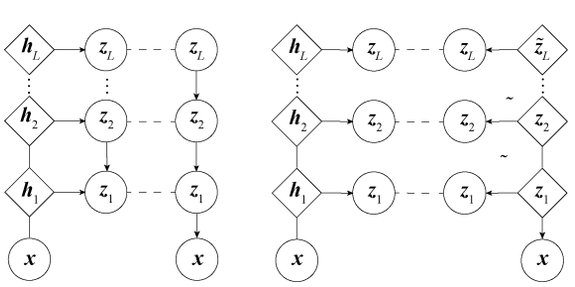
式中，$\boldsymbol{h}_l$表示$\boldsymbol{x}$自底而上的某个抽象级别，$S_l(\cdot)$则表示该抽象级别对应的抽象语义函数。

在基于VAE（Kingma和Welling，2022）的方法中，堆叠层级模型和架构层级模型的ELBO目标函数依次表示为

$$L_{\mathrm{ELBO}}=\sum_{l=0}^{L}E_{q(\boldsymbol{z},\boldsymbol{x})}\left[\log p(\boldsymbol{z}_l\mid\boldsymbol{z}_{>l})\right]+H\left(q(\boldsymbol{z}\mid\boldsymbol{x})\right)\tag{12}$$

$$L_{\mathrm{ELBO}}=E_{q(\boldsymbol{z}\mid\boldsymbol{x})}\left[\log p(\boldsymbol{x}\mid\boldsymbol{z})\right]-KL\left(q(\boldsymbol{z}\mid\boldsymbol{x})\mid p(\boldsymbol{z})\right)\tag{13}$$

式中，$H(\cdot)$代表一种分布的熵。



(a) 堆叠模型的推论模型和生成模型　　(b) 架构模型的推论模型和生成模型

图3　堆叠层级模型和架构层级模型的示意图

Fig. 3　Illustration of stacked hierarchical model and architectural hierarchical model ((a) inference and generative models for the stacked hierarchical model; (b) inference and generation models for the architectural hierarchical model)

2.3.2　特点及适用性

如图2（c）所示，层级解耦方式根据语义抽象水平将观测数据的生成因素$\boldsymbol{G}$解耦成L个不同层级的子集$\boldsymbol{G}^{(1)},\boldsymbol{G}^{(2)},\cdots,\boldsymbol{G}^{(L)}$，并用$L$个潜变量分别表示。层级越高的生成因素具有更加抽象、更高水平的语义信息。由于每个潜变量$\boldsymbol{z}_l(l=1,2,\cdots,L)$代表的是一个抽象水平的生成因素子集$\boldsymbol{G}^{(l)}(l=1,2,\cdots,L)$，而非单个具体的生成因素$g_i(i=1,2,\cdots,m)$，所以层级解耦是一种粗粒度的解耦表示学习。

基于VAE的层级解耦方法（Gulrajani等，2016；Hsu等，2017；Li等，2020g；Rezende等，2014；Sønderby等，2016；Zhao等，2017）用一个潜码$\boldsymbol{z}$作为高维数据$\boldsymbol{x}$的一种层级潜在表示，潜码$\boldsymbol{z}$被分解为不同的子块，每个子块代表一个抽象水平的潜变量。在类型上，基于VAE的层级解耦可细分为堆叠层级和架构层级两类。堆叠层级模型的代表工作之一是LVAE（ladder variational autoencoders）（Sønderby等，2016），采用多个VAE堆叠形成的梯形模型建模数据空间和潜空间的映射关系，通过自顶向下、连续依赖的生成分布和自底向上的近似后验分布表达层级化的解耦表示。架构层级模型的典型工作VLAE（variational ladder autoencoder）（Zhao等，2017）假设各潜变量独立分布，并通过单个VAE模型中不同深度的网络层捕获不同抽象水平的生成因素信息。基于其他生成模型的层级解耦方法（Li等，2020g；Peng等，2017；Singh等，2019；Tong等，2019）常采用多个潜码分别表达不同抽象水平的潜变量，通过分支化或阶段化的潜特征提取过程和数据生成过程实现层级解耦。

层级解耦方式适用于观测数据的生成因素存在一定联系的情况。此情况常见于真实数据中，因为真实数据的生成来源于生成因素复杂的交互网络中。例如，图像的对象与背景相关（鱼生活在水中，鸟在树上等），对象的外貌与对象的形状相关（鸟具有羽毛纹理，狗具有毛发纹理）。这些生成因素的联系语义抽象水平上的体现为，抽象水平高的生成因素决定了某些抽象水平低的生成因素分布。而层级

解耦不仅能将不同抽象水平的潜在变化因素分离，还能通过层级化的潜在表示捕获变化因素在不同抽象层面的相关性，提升生成图像的语义正确性和真实性。

2.4 非线性独立成分分析

2.4.1 问题描述

假设存在4个变量，分别为1个观测数据变量（随机向量）$\boldsymbol{x}\in\mathbf{R}^d$，1个低维的随机潜向量$\boldsymbol{z}\in\mathbf{R}^n(n\leqslant d)$，1个条件变量$\boldsymbol{u}\in\mathbf{R}^m$（比如类标签、时间序号或任何其他可以进一步观测的数据）和1个噪声变量$\boldsymbol{\varepsilon}\in\mathbf{R}^{d-n}$。潜变量$\boldsymbol{z}$和噪声变量$\boldsymbol{\varepsilon}$是不能被观测到的未知变量，它们共同组成了生成潜空间。而条件向量$\boldsymbol{u}$和数据点$\boldsymbol{x}$是可观测且通常已知。

Khemakhem等人（2020）提出潜变量模型的可识别性问题，即模型需要能保证，如果它学习到的边际似然性$p_\theta(\boldsymbol{x})$接近于观测数据$\boldsymbol{x}$的真实边际概率密度$p_{\theta'}(\boldsymbol{x})$，那么网络学习到的参数$\boldsymbol{\theta}$等于数据点$\boldsymbol{x}$的真实分布参数$\boldsymbol{\theta}'$，即

$$\forall(\boldsymbol{\theta},\boldsymbol{\theta}'):p_\theta(\boldsymbol{x})=p_{\theta'}(\boldsymbol{x})\Rightarrow\boldsymbol{\theta}=\boldsymbol{\theta}' \tag{14}$$

此时，模型学习到了潜变量正确的先验分布$p_\theta(\boldsymbol{z})=p_{\theta'}(\boldsymbol{z})$和正确的后验分布$p_\theta(\boldsymbol{z}\mid\boldsymbol{x})=p_{\theta'}(\boldsymbol{z}\mid\boldsymbol{x})$。即

$$\begin{aligned}p_\theta(\boldsymbol{x})&=\int p_\theta(\boldsymbol{x},\boldsymbol{z})\mathrm{d}\boldsymbol{z}\\p_\theta(\boldsymbol{x},\boldsymbol{z})&=p_\theta(\boldsymbol{x}\mid\boldsymbol{z})p_\theta(\boldsymbol{z})\end{aligned} \tag{15}$$

能够满足式（15）的模型就是可识别模型（identifiable model）。

在可识别潜变量模型中，条件生成模型定义为

$$p_\theta(\boldsymbol{x},\boldsymbol{z}\mid\boldsymbol{u})=p_f(\boldsymbol{x}\mid\boldsymbol{z})p_{T,\lambda}(\boldsymbol{z}\mid\boldsymbol{u}) \tag{16}$$

式中，T是统计量，λ是系数，f是一个由生成潜空间映射到数据空间的可逆确定性变化$\boldsymbol{x}=f(\boldsymbol{z},\boldsymbol{\varepsilon})$。潜变量的先验分布$p_{T,\lambda}(\boldsymbol{z}\mid\boldsymbol{u})$的概率密度为

$$p_{T,\lambda}(\boldsymbol{z}\mid\boldsymbol{u})=\prod_i\frac{Q_i(\boldsymbol{z}_i)}{Z_i(\boldsymbol{u})}\exp\left[\sum_{j=1}^k\boldsymbol{T}_{ij}(\boldsymbol{z}_i)\lambda_{ij}(\boldsymbol{u})\right] \tag{17}$$

式中，Q_i是一个基准测量方法，$Z_i(\boldsymbol{u})$是一个正则化常量，$\boldsymbol{T}_i=(T_{i,1},\cdots,T_{i,k})$是统计量，$\lambda_i(\boldsymbol{u})=(\lambda_{i,1}(\boldsymbol{u}),\cdots,\lambda_{i,k}(\boldsymbol{u}))$是严格依赖于条件变量$\boldsymbol{u}$的统计量对应参数。

通过最大化观测数据的对数似然性下界$L(\theta,\phi)$，其目标函数为

$$E_q\left[\log p_\theta(\boldsymbol{x}\mid\boldsymbol{u})\right]\geqslant L(\theta,\phi)=E_q\left[\log p_\theta(\boldsymbol{x},\boldsymbol{z}\mid\boldsymbol{u})\right]-\log q_\phi(\boldsymbol{z}\mid\boldsymbol{x},\boldsymbol{u}) \tag{18}$$

2.4.2 特点及适用性

如图2（d）所示，非线性独立成分分析（nonlinear ICA）理论认为，数据信号源$\boldsymbol{x}\in\mathbf{R}^d$是由多个独立的生成因素信号$\boldsymbol{z}\in\mathbf{R}^n$混合后，经过某种任意复杂的变换$\boldsymbol{f}$生成的，即$\boldsymbol{x}=f(\boldsymbol{z})$。生成因素的潜空间和观测空间之间存在的这种变换$f$是非线性可逆可微分的。非线性独立成分分析通过概率论和数理统计等知识，提供了一个严格的理论框架，根据一系列给定的数据点，用潜变量s模型f^{-1}恢复这些潜在的独立成分因素$\boldsymbol{z}=f^{-1}(\boldsymbol{z})$。其中，向量$\boldsymbol{z}$的每一个元素都分别表示一种独立的潜在生成因素，以实现各独立成分的完全分解，所以是一种细粒度的解耦表示学习。

基于非线性独立成分分析的解耦工作主要分为VAE和流模型两类。VAE方法的代表工作为Khemakhem等人（2020）提出的一种具有可识别性潜变量模型iVAE（identifiable variational autoencoder），其利用变分推论不断逼近潜变量的真实后验分布和观测数据的真实分布，并假设潜变量先验分布是依赖于一个额外的可观测条件变量的因子化条件先验实现了解耦。基于流模型的nonlinear ICA工作（Li等，2020c；Sorrenson等，2020）通过一系列可逆变换函数构建观测数据空间和潜空间之间任意复杂的概率密度变换，以直接最大化观测数据的边际似然为目标函数，分离出潜在生成因素并学习到了潜变量和观测数据的真实分布。

由于非线性独立成分分析中关于模型可识别性的数学理论是基于观测数据的本质生成因素已知的假设，所以此类解耦表示学习方法适用于生成因素已知的模拟数据和一些简单低维的合成数据集，如随机生成的2维数据点和EMNIST（extended mixed national institute of standards and technology database）（Cohen等，2017）。对于任意复杂的真实世界数据，想要获取观测数据的本质生成因素的先验知识是困难的，因此非线性独立成分分析方法距离高维复杂的真实世界数据的解耦应用还具有一定的距离。

2.5 解耦表示学习类型对比与总结

本节根据潜变量之间的相关性或独立性，将解耦表示学习分为维度独立解耦、语义独立解耦、层级

解耦和非线性独立成分分析4类。对于每类解耦方式，从问题描述、潜变量与生成因素的对应关系、生成模型及其代表作、适用范围4个角度进行了分析。维度独立解耦基于生成因素间相互独立的假设，将观测数据的生成因素逐个分离并映射到潜向量的各维度，适用于学习简单合成数据的解耦表示。语义独立解耦基于某些语义信息相互独立的假设，针对观测数据的特性，将生成因素按特定语义逐群解耦并映射到不同的潜空间适用于真实复杂数据的解耦表示和实际问题应用。层级解耦基于不同抽象水平的生成因素之间存在相关性的假设，将生成因素自底向上逐群解耦并映射至不同语义抽象水平的潜空间，形成层级化的解耦表示。非线性独立成分分析提供了一个可识别的方法，通过一个非线性可逆生成器分解观测数据中混合的潜在生成因素。表1总结了每种类型解耦表示方法的特点并横向对比各类方法。表2列举了每类解耦表示方法的代表性工作，并总结了其研究贡献。

表1　4种解耦表示学习类型的对比

Table 1　Comparison of four DRL types

对比属性	维度解耦	语义解耦	层次解耦	非线性独立成分分析
生成因素间关系	互相独立	逐群独立	逐群相关	互相独立
潜码个数	1个	2个或多个	2个或多个	1个
潜空间维度	1维向量	1维向量或3维矩阵	1维向量	1维向量
潜码与潜变量的对应关系	1对N	1对1	1对1	1对N
解耦粒度	细粒度	粗粒度	粗粒度	细粒度
解耦语义可控性	不可控	可控	可控	不可控
模型可识别性	否	否	否	是
适用范围	较为复杂的合成数据	真实世界数据	真实世界数据	简单的合成数据

表2　4种解耦表示学习类型的代表论文总结

Table 2　Summary of representative papers on four types of DRL

解耦类型	模型名称	发表会议	主要贡献
维度解耦	InfoGAN（Chen等，2016）	NIPS 16	通过最大化观测数据和潜码之间的互信息学习鼓励解耦表达。
	β-VAE（Higgins等，2017）	ICLR 17	引入一个大于1的系数β惩罚KL散度项，约束潜变量z的信息瓶颈。
	Annealed VAE（Burgess等，2018）	NIPS Workshop 17	在KL散度项内引入变量c作为目标，进一步约束潜信息编码容量。
	β-TCVAE（Chen等，2019）	NIPS 18	分解KL散度项为互信息、TC和逐维度KL散度3个子项，并分别用α,β,γ共3个系数进行惩罚。
	FactorVAE（Kim和Mnih，2018）	ICML 18	在传统VAE的ELBO末尾添加一个TC惩罚项，促进潜向量z的各维度尽可能独立。
	DIP-VAE（Kumar等，2018）	ICLR 18	在传统VAE的ELBO末尾添加一个解耦正则项，约束潜向量z的聚合后验分布和先验分布的协方差相近。
	ID-GAN（Lee等，2020）	ECCV 20	使用VAE模型学习解耦表示c并蒸馏掉无关信息s，GAN模型的对抗学习生成高质量生成图。
语义解耦	DRNET（Denton和Birodkar，2017）	NIPS 17	将每个视频帧分解成静止不变分量和短时变化分量。静止分量捕获视频帧跨时间共享的信息，变化分量捕获对象在各视频帧中变化的动作和位置信息。
	DSVAE（Li和Mandt，2018）	ICML 18	将序列解耦为时间不变性和时间依赖性潜变量，应用LSTM（long short-term memory）（Hochreiter和Schmidhuber，1997）序列先验保持生成序列的顺序一致性。
	DDPAE（Hsieh等，2018）	NIPS 18	以无监督学习的方式将视频数据的表示解耦为内容分量和姿势分量，用以视频生成和视频预测。

续表2　4种解耦表示学习类型的代表论文总结

Table 2　Summary of representative papers on four types of DRL (continued)

解耦类型	模型名称	发表会议	主要贡献
语义解耦	MoCoGAN(Tulyakov等,2018)	CVPR 18	将视频帧分解为内容潜变量和动作潜变量,引入两个判别器分别评估生成帧的图像和动作的真实性。
	DRIT(Lee等,2018)	ECCV 18	将图像嵌入到域共享一致的内容空间和域特有的属性空间,并采用交叉循环一致性损失函数促进解耦。
	Gonzalez等人(2018)	NIPS 18	将图像表示分解为域共享因素、域专属因素,通过跨域自编码器实现表示解耦和跨域图像翻译。
	UFDN(Liu等,2018a)	NIPS 18	提出一个统一特征解耦网络UFDN,将N个域的图像表示为域不变性表征z和域特殊性向量$\boldsymbol{V}_c$,以实现连续的跨域图像翻译和图像编辑等操作。
	D2AE(Liu等,2018e)	CVPR 18	将人脸潜在表示解耦为身份蒸馏特征和身份去除特征,采用双流设计网络对抗学习这个互相独立且互补的特征。
	CDRD(Zhu等,2017)	CVPR 18	提出一种跨域解耦表示学习模型将空间分解成源域空间、目标域空间和公共空间,并认为源域空间和目标域空间存在一部分重叠空间,即为公共空间。
	DMIT(Yu等,2019)	NIPS 19	将图像解耦到内容潜空间和风格潜空间,通过操作不同的潜变量同时实现图像多模态生成和跨域翻译。
	GaitNet(Li等,2020e)	CVPR 19	将行人图像的表示分解为姿态和外貌特征,融合连续时间内的多个姿态特征生成步态特征。
	MixNMatch(Li等,2020f)	CVPR 20	将图像解耦为背景、物体姿势、形状和纹理等潜变量,通过潜变量混合重组实现条件图像生成。
	S3VAE(Zhu等,2020b)	CVPR 20	利用自监督信号和一系列辅助任务促进解耦学习,将序列的表示分解为静态因素和动态因素。
层级解耦	FHVAE(Hsu等,2017)	NIPS 17	提出因子化层次VAE,对不同的潜变量集应用序列相关的先验和序列无关的先验。
	VLAE(Zhao等,2017)	ICML 17	提出变分梯形自编码器,用模型中不同深度的网络层学习各个抽象水平的潜变量,获得层级化的表示。
	Peng等人(2017)	ICCV 17	采用3D可变形模型将人脸图像嵌入低维的潜空间并将其分解为身份特征空间和非身份特征空间,后者进一步被层级分解为面部关键点特征空间和姿势特征空间。
	FineGAN(Singh等,2019)	ICCV 19	将图像解耦为姿势、背景、形状和外貌4个不同抽象水平的潜变量,经过3个阶段层级化地生成图像。
	proVLAE(Li等,2020g)	ICLR 20	通过渐进式增长的网络依次学习抽象水平从高到低的潜变量,形成数据的层级解耦表示。
非线性独立成分分析	DUAL-GLOW(Sun等,2019a)	ICCV 19	提出了一个基于流的生成模型DUAL-GLOW,通过将医疗图像中附加信息(年龄、脑状态等)从本源信息中分离,实现不同图像域间的模态迁移和条件图像生成。
	FUNS(Kondo等,2019)	NIPS 19	提出了一个基于流的图像翻译模型FUNS,既可将目标图像的多样性分解为条件依赖和条件不依赖部分实现特征分解,也可从条件图像中生成高质量多样性图像。
	iVAE(Khemakhem等,2020)	AISTATS 20	提供了一种将VAE生成模型和nonlinear ICA相结合的通用性网络框架iVAE,并提出了潜变量和观测变量分布的可识别性概念,对该性质进行了数学描述和理论推导。
	GIN(Sorrenson等,2020)	ICLR 20	改进了realNVP流模型(Dinh等,2017),提出了一种新网络GIN,进一步完善了iVAE的理论,将模型可识别性推广到真实世界数据集中其本质生成因素未知的情况。
	Klindt等人(2021)	ICLR 21	提供了自然数据的生成因素随着时间的变化是稀疏分布的证明,并基于稀疏时域迁移特性,用拉普拉斯稀疏先验分布代替传统的高斯分布族,提升了模型的解耦性能。

3 损失函数

解耦表示学习通过损失函数的约束，鼓励潜在表示满足模块化和显式性的解耦表征特性。本节根据损失函数想要实现的解耦表征特性，将常用损失函数分为模块化约束、显式性约束和多目的约束3类，并依次进行介绍。

3.1 模块化约束

1）交叉重建损失（cross-reconstruction loss）。在语义独立解耦方式的某些情况下，观测数据的潜在表示被分解为共享分量 $\boldsymbol{C}$ 和特有分量 $\boldsymbol{s}$（例如Gonzalez-Garcia等人（2018）将跨域图像解耦为域共享内容特征和域特有风格特征）。对于具有一部分相同语义信息的数据对 $\{\boldsymbol{x},\boldsymbol{y}\}$，共享分量编码器 $\{E_x^c,E_y^c\}$ 和特有分量编码器 $\{E_x^s,E_y^s\}$ 将其嵌入到共享的公共空间 $\boldsymbol{c}$ 和特定的独有空间 $\boldsymbol{S}_X$ 和 $\boldsymbol{S}_Y$，具体为

$$\begin{aligned}\{\boldsymbol{c}^x,\boldsymbol{s}^x\}&=\{E_x^c(\boldsymbol{x}),E_x^s(\boldsymbol{x})\}\quad c^x\in\boldsymbol{C},\boldsymbol{s}^x\in\boldsymbol{S}_X\\\{\boldsymbol{c}^y,\boldsymbol{s}^y\}&=\{E_y^c(\boldsymbol{y}),E_y^s(\boldsymbol{y})\}\quad c^y\in\boldsymbol{C},\boldsymbol{s}^y\in\boldsymbol{S}_Y\end{aligned}\tag{19}$$

式中，$\{\boldsymbol{c}^x,\boldsymbol{s}^x\}$ 分别表示 $\boldsymbol{x}$ 的共享分量和特有分量，$\{\boldsymbol{c}^y,\boldsymbol{s}^y\}$ 同理。理想情况下，共享分量 $\boldsymbol{c}^x$ 和 $\boldsymbol{c}^y$ 是相同一致的。为了鼓励 E^c 将 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 之间的共享信息编码进公共空间 $\boldsymbol{C}$，而 E^s 将剩余的特有信息映射到 $\boldsymbol{S}_X$ 和 $\boldsymbol{S}_Y$，提出交叉重建损失函数，即交换 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 的共享分量，重组的潜在表示经过生成器 $\{G_x,G_y\}$ 生成输入的重建，用 L_1 重建损失函数度量，具体为

$$\begin{aligned}L_{\text{crec}}=&E_{x,y}\left[\left\|G_x(\boldsymbol{c}^y,\boldsymbol{s}^x)-x_2^2\right\|\right]+\\&E_{x,y}\left[\left\|G_x(c^x,\boldsymbol{s}^y)-y_2^2\right\|\right]\end{aligned}\tag{20}$$

2）互信息最小化（mutual information minimization）。在语义独立解耦方式中，假设观测数据 $\boldsymbol{x}$ 被解耦为两个互相独立的特征 $\{\boldsymbol{u},\boldsymbol{v}\}$。为了更好地解耦，引入互信息（mutual information，MI）衡量两个潜变量间的相互依赖性。通过最小化特征 $\{\boldsymbol{u},\boldsymbol{v}\}$ 的互信息，鼓励这两个潜变量的信息是相互排斥的。互信息定义为联合分布与潜变量边际分布乘积的KL散度，具体为

$$\begin{aligned}L_{\text{MI}}(\boldsymbol{u},\boldsymbol{v})=&KL(q(\boldsymbol{u},\boldsymbol{v})\|q(\boldsymbol{u})q(\boldsymbol{v}))=\\&[H(\boldsymbol{u})+H(\boldsymbol{v})-H(\boldsymbol{u},\boldsymbol{v})]\end{aligned}\tag{21}$$

式中，$H(\cdot)=-E_{q(\cdot)}[\log q(\cdot)]$ 代表潜变量边际分布或联合分布的熵。

3.2 明确性约束

1）自重建损失（self-reconstruction loss）。为保证提取的潜在表示 $\boldsymbol{z}$ 完整编码了输入数据的所有生成因素信息，潜码 $\boldsymbol{z}$ 经过生成器 G 生成的重建 $\tilde{\boldsymbol{x}}=G(\boldsymbol{z})$ 应尽可能接近原输入 $\boldsymbol{x}$，两者之间的差异用 L_1 距离衡量，即

$$L_{\text{rec}}=\left|\boldsymbol{x}-G(\boldsymbol{z})\right|_1\tag{22}$$

2）互信息最大化（mutual information maximization）。互信息最大化由Chen等人（2016）在解耦表示模型InfoGAN（information-theoretic extension to the generative adversarial network）中首次提出，将观测数据 $\boldsymbol{x}$ 解耦表示为一个代表噪声的随机向量 $\boldsymbol{z}$ 和一个捕获显著语义信息的潜码 $\boldsymbol{c}$。在互信息理论中，如果 A 和 B 之间存在一种确定可逆的关联性，那么两者的互信息应该达到最大值。因此通过最大化潜码 $\boldsymbol{c}$ 和生成数据 $G(\boldsymbol{z},\boldsymbol{c})$ 之间的互信息，增强潜码的可解码性，使潜码中的信息在生成过程中不被丢失。互信息 I 可以被表达为两个分布的熵 $H(\cdot)$ 差值，具体为

$$-I(\boldsymbol{c};G(\boldsymbol{z},\boldsymbol{c}))=-[H(\boldsymbol{c})-H(\boldsymbol{c}|G(\boldsymbol{z},\boldsymbol{c}))]\tag{23}$$

由于直接最大化互信息所需的后验分布 $P(\boldsymbol{c}|\boldsymbol{x})$ 难以求解，常用解码器参数化实现一个辅助分布 $Q(\boldsymbol{c}|\boldsymbol{x})$ 近似 $P(\boldsymbol{c}|\boldsymbol{x})$，进而最大化互信息的变分下界，具体为

$$\begin{aligned}&I(\boldsymbol{c};G(\boldsymbol{z},\boldsymbol{c}))=H(\boldsymbol{c})-H(\boldsymbol{c}|G(\boldsymbol{z},\boldsymbol{c}))=\\&E_{x\sim G(z,c)}\left[E_{c'\sim P(c|x)}\left[\log P(\boldsymbol{c}'|\boldsymbol{x})\right]\right]+H(\boldsymbol{c})\times\\&E_{x\sim G(z,c)}\left[\underbrace{D_{KL}\left(P(\cdot|\boldsymbol{x})\|Q(\cdot|\boldsymbol{x})\right)}_{\geqslant 0}+E_{c'\sim P(c|x)}\left[\log Q(\boldsymbol{c}'|\boldsymbol{x})\right]\right]+\\&H(\boldsymbol{c})\times E_{x\sim G(z,c)}\left[E_{c'\sim P(c|x)}\left[\log Q(\boldsymbol{c}'|\boldsymbol{x})\right]\right]+H(\boldsymbol{c})\end{aligned}\tag{24}$$

3）三元组损失（triplet loss）。在语义独立解耦表示中，假设观测数据 $\boldsymbol{x}$ 的变化因素可分为互相独立的两类，分别具有物理语义 s_1 和 s_2，用潜码 $\{\boldsymbol{u},\boldsymbol{v}\}$ 表达。为了确保潜码仅编码了对应的变化因素信息，引入三元组损失并构造三元组 $\{a,p,n\}$ 训练数据。其中 a 为基准样本（anchor），p 代表在语义 s_1 上和 a 相近的正样本（positive），n 则代表在语义 s_1 上和 a 不相近的负样本（negative）。因此，a 和 p 的特征 $\boldsymbol{u}$ 之间的距离应该小于 a 和 n 的特征 $\boldsymbol{u}$ 之间的距离，具体为

$$L_{\text{triplet}}=\max\left(\left\|\boldsymbol{u}^a-\boldsymbol{u}^p\right\|^2-\left\|\boldsymbol{u}^a-\boldsymbol{u}^n\right\|^2+m,0\right)\tag{25}$$

式中，m是一个控制正负样本的距离的阈值，实际值需要手动设置。

4）潜码相似性（latent similarity）。与交叉重建损失函数的应用背景相似，假设观测数据$\boldsymbol{x}$的潜在表示被分解为不变性分量$\boldsymbol{c}$和变化性分量$\boldsymbol{s}$，对于具有一部分相同不变性信息的数据对$\{\boldsymbol{x},\boldsymbol{y}\}$，为了确保不变性特征编码器$E^c$只捕获观测数据中不变的静态生成因素，$\boldsymbol{x}$和$\boldsymbol{y}$的不变性分量$\boldsymbol{c}$应该接近，用$L_2$损失惩罚两者的误差，具体为

$$L_{\text{similarity}}\left(E^c\right)=\left\|E^c(\boldsymbol{x})-E^c\left(\boldsymbol{y}\right)\right\|_2^2 \tag{26}$$

5）交叉熵损失（cross entropy loss）。为了鼓励潜码有效编码了对应的生成因素信息，常利用分类或者识别辅助任务。潜码$\boldsymbol{z}$通过一个分类器C预测出对应生成因素的真实值或者标签$\boldsymbol{y}$，采用交叉熵损失函数训练，具体为

$$L_{\text{CE}}=\sum_{\boldsymbol{c}} I(\boldsymbol{c}=\boldsymbol{y})\log p\left(\boldsymbol{c}\mid\boldsymbol{z}\right) \tag{27}$$

式中，$I(\boldsymbol{c}=\boldsymbol{y})$是二元指标函数，如果输出的预测值$\boldsymbol{c}$是正确的分类标签值$\boldsymbol{y}$，则为0，否则为1。

3.3 多目的性约束

1）循环一致性损失（cycle consistency loss）。Zhu等人（2018）在CycleGAN（cycle-consistent generative adversarial network）模型中首次提出循环一致性损失。为了同时满足模块化和显式性两个解耦表征属性，在无监督解耦学习工作中（Eom和Ham，2019；Lee等，2018；Zhang等，2019a；Zhu等，2018）广泛应用。循环一致性包含两个阶段的生成过程。假设将观测数据解耦为两个互补且互相独立的特征$\boldsymbol{u}$和$\boldsymbol{v}$。在前向生成阶段中，给定一个没有对应关系的图像对$\boldsymbol{x}$和$\boldsymbol{y}$，将它们编码为$\{\boldsymbol{u}^x,\boldsymbol{v}^x\}$和$\{\boldsymbol{u}^y,\boldsymbol{v}^y\}$。通过交换特征$\boldsymbol{u}^x$和$\boldsymbol{u}^y$，生成$\{\tilde{\boldsymbol{x}},\tilde{\boldsymbol{y}}\}$，具体为

$$\tilde{\boldsymbol{x}}=\boldsymbol{G}_x\left(\boldsymbol{u}^y,\boldsymbol{v}^x\right),\tilde{\boldsymbol{y}}=\boldsymbol{G}_y\left(\boldsymbol{u}^x,\boldsymbol{v}^y\right) \tag{28}$$

在反向生成阶段中，将$\tilde{\boldsymbol{x}}$和$\tilde{\boldsymbol{y}}$编码为$\{\boldsymbol{u}^{\tilde{x}},\boldsymbol{v}^{\tilde{x}}\}$和$\{\boldsymbol{u}^{\tilde{y}},\boldsymbol{v}^{\tilde{y}}\}$。通过再次交换特征$\boldsymbol{u}$，生成$\{\hat{\boldsymbol{x}},\hat{\boldsymbol{y}}\}$，具体为

$$\hat{\boldsymbol{x}}=\boldsymbol{G}_x\left(\boldsymbol{u}^{\tilde{y}},\boldsymbol{v}^{\tilde{x}}\right),\hat{\boldsymbol{y}}=\boldsymbol{G}_y\left(\boldsymbol{u}^{\tilde{x}},\boldsymbol{v}^{\tilde{y}}\right) \tag{29}$$

经过两次编码—交换—解码过程，$\{\hat{\boldsymbol{x}},\hat{\boldsymbol{y}}\}$应该重建原图像$\boldsymbol{x}$和$\boldsymbol{y}$。为了强制执行此约束，将循环一致性损失表达为

$$L_{\text{cyc}}=\left\|\hat{\boldsymbol{x}}-\boldsymbol{x}\right\|_1+\left\|\hat{\boldsymbol{y}}-\boldsymbol{y}\right\|_1 \tag{30}$$

2）潜码回归损失（latent regression loss）。在语义解耦方式中，为了鼓励潜码的信息在生成数据中被有效解码表达，且各解耦特征彼此分离，常使用潜码回归函数。假设观测数据$\boldsymbol{x}$被分解为特征$\{\boldsymbol{u},\boldsymbol{v}\}$，而特征$\boldsymbol{v}$服从先验分布$N(0,I)$。从先验高斯分布中随机采样获得一个潜向量$\boldsymbol{z}$，并从输入数据$\boldsymbol{x}$中提取特征$\boldsymbol{u}$，输入生成器$G$得到生成样本。

生成的样本再送入编码器解耦提取特征，则应该重建出潜码$\boldsymbol{u}$和潜向量$\boldsymbol{z}$，重建损失用L_1距离度量，具体为

$$\begin{aligned}L_{\text{reg}}=&E_{\substack{\boldsymbol{x}\sim X\\ \boldsymbol{z}\sim N(0,I)}}\left[\left\|E^u\left(G\left(E^u(\boldsymbol{x}),\boldsymbol{z}\right)\right)-E^u(\boldsymbol{x})\right\|_2^2\right]+\\&E_{\substack{\boldsymbol{x}\sim X\\ \boldsymbol{z}\sim N(0,I)}}\left[\left\|E^v\left(G\left(E^u(\boldsymbol{x}),\boldsymbol{z}\right)\right)-\boldsymbol{z}\right\|_2^2\right]\end{aligned} \tag{31}$$

3）变分下界（evidence lower bound）。在VAE（Kingma和Welling，2022）的解耦表示模型中，常采用变分下界或其变体作为损失项。为鼓励潜码捕获所有生成因素信息且促进潜变量间互相独立，最小化负变分下界（negative variational lower bound），计算表达式为

$$L_{\text{VAE}}=E_{q(\boldsymbol{z}\mid\boldsymbol{x})}\left[-\log p\left(\boldsymbol{x}\mid\boldsymbol{z}\right)\right]+KL\left(q\left(\boldsymbol{z}\mid\boldsymbol{x}\right)\mid p\left(\boldsymbol{z}\right)\right) \tag{32}$$

式中，右边第1项$E_{q(\boldsymbol{z}\mid\boldsymbol{x})}\left[-\log p\left(\boldsymbol{x}\mid\boldsymbol{z}\right)\right]$可以看做是重建损失项，约束潜码信息经过解码后可以复原出输入数据。第2项是KL散度，约束推论模型参数化的潜码后验分布与$q(\boldsymbol{z}\mid\boldsymbol{x})$假设的先验分布$p(\boldsymbol{z})$接近。

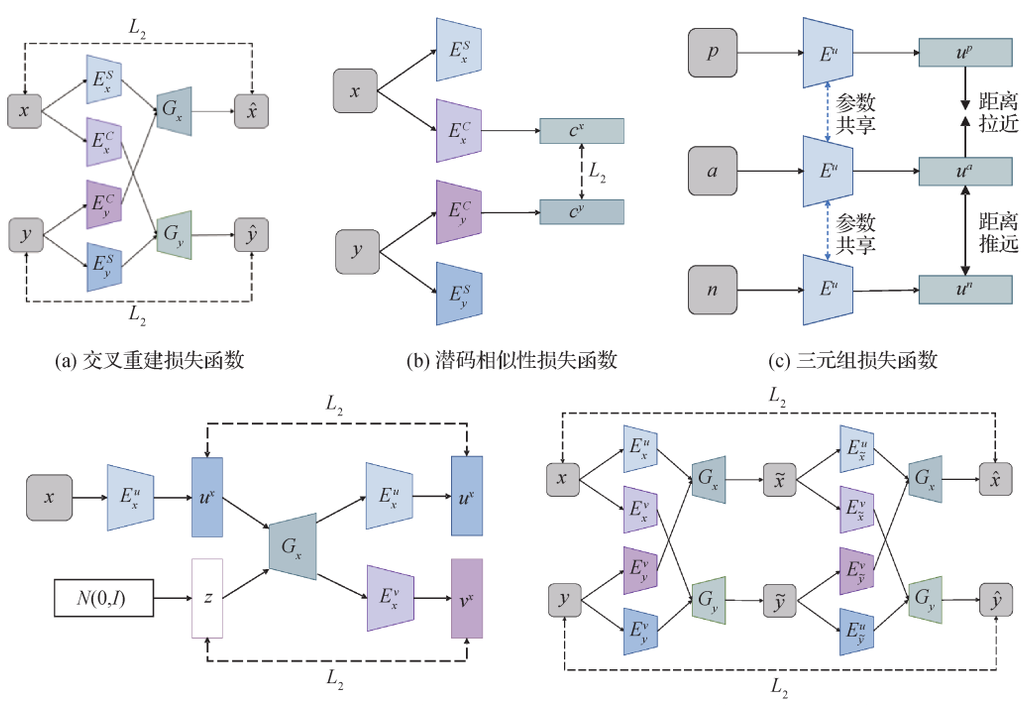
3.4 小结

从损失函数的动机出发，根据其想要实现的解耦表示性质，将解耦表示学习中常用的损失函数归为3类。1）模块化约束。约束解耦表示中的单个潜变量只捕获单个或单组变化因素、促进变化因素之间互相分离，以实现模块化性质。2）明确性约束。鼓励潜在表示的当个潜变量有效编码了对应生成因素的真实值，整个潜在表示则完整包含了所有生成因素的信息，以实现明确性性质。3）多目的性约束。能同时促进解耦表示模块化、紧凑性、明确性等多个解耦表示性质的损失项。在大多数解耦表示工作中，模型通常会组合上述归纳中的多个损失约束项形成最终的混合目标函数。表3进一步对各类损失函数的适用范围和局限性进行对比。图4为几种解耦表示学习中常用的损失函数示意图，阐释了解耦表示学习中几种常用的损失函数。表4总结了使用混合目标函数的经典解耦表示工作。

表 3　解耦表示学习中常用的损失函数对比

Table 3　Comparison of commonly used loss functions in DRL

损失函数类型	损失函数	适用场景	局限性
模块化约束	交叉重建损失函数	语义解耦表示学习方式中，分解的语义分量在多个样本中具有一致性或是共享的。	L_1 或 L_2 损失函数容易使模型输出的重建图像模糊，且要求严格对齐的成对数据样本。
	互信息最小化	观测数据的生成因素间是独立的，不能存在任何依赖关系。	不适用于层级解耦表示学习方式。
明确性约束	自重建损失函数	要求模型学习到的解耦表示中需要具备完备性这一特性。	L_1 或 L_2 损失函数容易让模型输出的重建图像模糊，仅限于生成类下游问题。
	互信息最大化	适用于维度解耦表示学习方式。	直接最大化互信息所需的潜变量后验分布难以求解，需要辅助分布来近似，易引入误差。
	三元组损失	观测数据需具有类标签等附加信息用以根据锚点样本确定正负样本。	需要以在线和离线方式构造特定的三元组样本作为训练数据，且三元组样本的选择在一定程度上会影响模型的训练效果。
	潜码相似性	语义解耦表示学习方式中，分解的语义分量在多个样本中具有一致性或是共享的。	容易使所有特征信息泄漏到另一个特征分量中，使得该损失函数约束的分量趋近于一个接近于零的无意义常量。
	交叉熵损失	观测数据具有类标签或下游任务为识别分类问题。	训练需要人工标注标签这样的显式信息，不适用于无监督学习。
多目的性约束	循环一致性	观测数据跨域或下游任务为域迁移、域适应问题。	只适用于观测存在多模态跨域的样本情况，对单域观测数据不适用。
	潜码回归损失函数	跨域图像翻译、风格迁移等下游问题。	随机采样中包含潜变量的分布服从标准正态分布的假设具有一定的局限性。
	变分下界	潜变量生成模型为 VAE 模型或流模型。	由于 KL 散度中潜变量的真实后验分布不可求解，所以通过变分推论近似逼近，容易使模型在训练中陷入较差的次优解。



(a) 交叉重建损失函数　(b) 潜码相似性损失函数　(c) 三元组损失函数

(d) 循环一致性损失函数　(e) 潜码回归损失函数

图 4　几种解耦表示学习中常用的损失函数示意图

Fig. 4　Illustration of several loss functions commonly used in DRL((a)cross-reconstruction loss; (b)latent similarity loss; (c)triplet loss; (d)cycle consistency loss; (e)latent regression loss)

表4　采用混合损失函数的解耦表示学习模型的论文总结

Table 4　The summary of papers on DRL models using hybrid loss functions

模型	交叉重建	自重建	互信息	三元组	潜码相似性	交叉熵	循环一致性	潜码回归	变分下界
Peng等人(2017)	√	√	–	–	–	√	–	–	–
FHVAE(Hsu等,2017)	–	–	–	–	–	√	–	–	√
DRNET(Denton和Birodkar,2017)	–	–	–	–	√	√	–	–	–
UFDN(Liu等,2018a)	–	–	–	–	–	√	–	–	√
DRIT(Lee等,2018)	–	√	–	–	–	–	√	√	–
Gonzalez-Garcia等人(2018)	√	–	–	–	√	–	–	√	–
MIX(Zwicker等,2018)	–	√	–	–	–	√	–	–	–
CDRD(Zhu等,2017)	–	–	–	–	–	√	–	–	√
Liu等人(2018c)	–	√	–	–	–	√	–	–	–
D^2AE(Liu等,2018e)	–	√	–	–	–	√	–	–	–
DMIT(Yu等,2019)	–	–	–	–	–	–	–	√	√
Zhang等人(2019a)	√	–	–	–	√	√	–	–	–
Kotovenko等人(2019)	–	√	–	√	–		–	√	–
DADA(Peng等,2019)	–	–	√	–	–	√	–	–	–
Zhang等人(2019b)	–	–	–	–	–	√	√	–	–
Hamaguchi等人(2019)	–	–	–	–	√	√	–	–	√
IS-GAN(Eom和Ham,2019)	√	√	–	–	–	√	–	–	–
Guo等人(2019)	√	–	√	√	–	–	–	–	–
S3VAE(Zhu等,2020b)	–	–	√	√	–	√	–	–	√
IIAE(Hwang等,2020)	–	–	√	–	–	–	–	–	√
Sanchez等人(2020)	–	–	√	–	√	–	–	–	–
ID-GAN(Lee等,2020)	–	–	√	–	–	–	–	–	√
Li等人(2020f)	–	√	–	–	√	–	–	√	–
ICAM(Bass等,2020)	–	√	–	–	–	√	√	√	–
DG-Net++(Zou等,2020)	–	–	–	–	–	√	√	–	–
DG-VAE(Pu等,2020)	√	√	–	√	–	√	√	–	–

注:“√”代表该解耦工作的混合目标函数包含了相应的损失函数项,“–”代表不包含。

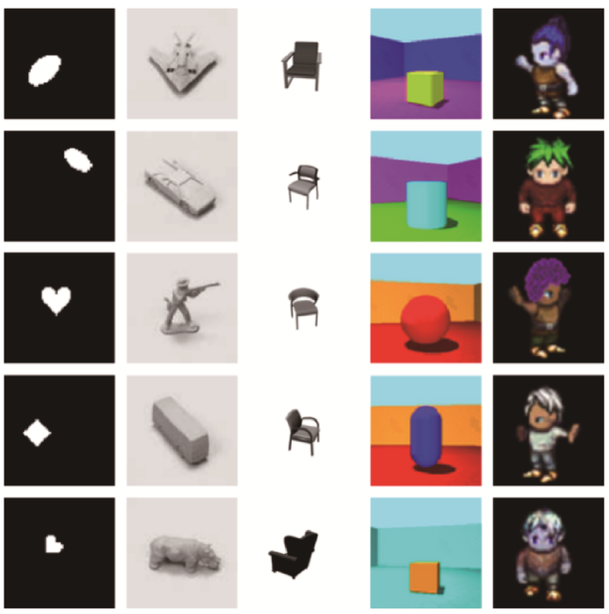
4　数据集与评估指标

4.1　解耦对象与数据集

本节总结了解耦表示学习中一些常用的图像、视频和音频等多媒体数据集。在第1节中提到,大多数解耦工作是有监督或弱监督学习方式,需要观测数据的部分或完整的生成因素真实值作为监督信号。由于真实世界数据的产生成本很高,且有时难以获取生成因素的真实值,许多解耦表示工作(Chen等,2019; Higgins等,2017; Hsieh等,2018; Li和Mandt,2018)都在简单的合成数据集上训练和测试。如图5所示,解耦评估基准数据集均为合成数据集。表5总结了解耦学习工作常用视觉数据集及其数据容量、分辨率和变化因素个数等具体信息。

4.1.1　图像数据集

1) dSprites数据集。dSprites(disentanglement testing sprites dataset)(2D Shapes)(Matthey等,2017)



(a) dSprites (b) smallNORB (c) 3D Chairs (d) 3D Shapes (e) Sprites

图5　几种解耦测评基准数据集

Fig. 5　Several DRL evaluation benchmarks ((a) dSprites; (b) smallNORB; (c) 3D Chairs; (d) 3D Shapes; (e) Sprites)

是一个包含737 280幅2维图像的合成数据集，每个2维图像是64 × 64像素的黑白图像，由6个具有真实值(ground truth)的独立的潜在因素生成，6个潜在因素是颜色、形状、大小、旋转角度、x位置和y位置。某些研究工作(Lee等，2020；Locatell等，2019a)中使用其变种数据集color-dSprites，noisy-dSprites，scream-dSprites。

2) MNIST数据集。MNIST(mixed national institute of standards and technology database)(LeCun等，1998)是一个由70 000幅28 × 28像素的黑白图像组成的手写数字数据集，其中60 000幅图像属于训练集，其余图像用于测试。

3) Fashion-MNIST数据集。Fashion-MNIST(Xiao等，2017)包含10种类别的70 000幅不同商品的正面图像，由60 000幅训练样本和10 000幅测试样本组成，每幅图像是28 × 28像素的灰度图。

4) smallNORB数据集。smallNORB (small NORB dataset)(LeCun等，2004)是一个3维物体的图像数据集，包含四脚动物、人形、飞机、卡车和汽车5个类别的50个玩具，每个物体6个照明条件、9个仰角和18个方位角的情况下由两个摄像头成像为92 × 92像素的照片。

5) 3D Shapes数据集。3D Shapes(Burgess和Kim，2018)是一个3维图像的合成数据集，包含48 000幅64 × 64像素的RGB图像和48 000个标签。每幅3维图像由6个具有真实值的独立的潜在因素生成，这6个潜在因素是地板颜色、墙壁颜色、物体颜色、物体大小、形状和视角方位。

6) 3D Chairs数据集。3D Chairs(Aubry等，2014)是一个由1 357把不同椅子的3维CAD模型渲染的合成数据集。每个CAD模型都有真实椅子的纹理，用60个不同姿势渲染。每幅渲染图为150×150像素的灰度图像。

7) 3D Faces数据集。3D Faces(Paysan等，2009)是一个使用Paysan等人(2009)的3维人脸模型渲染的合成数据集，包含3个变化属性，即脸的身份属性(形状/纹理)、姿态和光线，每幅图像为150 ×150像素的灰度图。

8) 3D Cars数据集。3D Cars(Fidler等，2012)是一个合成的3维渲染图像数据集，使用199个CAD模型，分别从24个旋转角度生成64 × 64像素的彩色渲染静止图。

9) SVHN数据集。SVHN(street view house numbers)(Netzer等，2011)是一个从谷歌街景图像的门牌中获得的数字和编号数据集，由超过600 000幅图像组成。裁剪格式的数据集中，图像分辨率为32×32像素。

10) CelebA数据集。CelebA(large-scale celebfaces attributes)(Liu等，2015)是一个大规模的人脸属性数据集，包含10 177个名人身份的202 599幅人脸图像，每幅图像有40个属性注释。

11) CUB-200-2011数据集。CUB-200-2011 (caltech-UCSD birds 200)(Wah等，2011)是一个真实图像数据集，包含200种鸟类的11 788幅图像，每幅图像的注释包含1个边界框、15个部位关键点和312个属性。

4.1.2　视频数据集

1) Bouncing balls数据集。Bouncing balls(Chang等，2017)是一个合成的视频序列数据，模拟3个小球在1个2维盒子里移动，每个球的初始位置、速度和质量是随机设定的，其中每个视频帧是一幅32×32像素的黑白图像。

2) Sprites数据集。Sprites(Li和Mandt，2018)是卡通人物视频序列数据集。人物的外貌包含肤色、上衣、裤子、发型4个属性，每个外貌属性有6个可能

表5　解耦表示学习常用的视觉数据集总结

Table 5　The summary of commonly used visual datasets in DRL

数据类型	数据集	数据总量	分辨率	变化因素/个	合成数据集	解耦测评基准数据集
图像	dSprites(Matthey等,2017)	737 280幅	64×64	6	√	√
	MNIST(LeCun等,1998)	70 000幅	28×28	–	×	×
	Fashion-MNIST(Xiao等,2017)	70 000幅	28×28	–	×	×
	smallNORB(LeCun等,2004)	48 600幅	92×92	5	√	√
	3D Shapes(Burgess和Kim,2018)	48 000幅	64×64	6	√	√
	3D Chairs(Aubry等,2014)	81 420幅	150×150	3	√	√
	3D Faces(Paysan等,2009)	239 840幅	150×150	3	√	√
	3D Cars(Fidler等,2012)	17 568幅	64×64	3	√	√
	SVHN(Netzer等,2011)	77 163幅	32×32	–	×	×
	CelebA(Liu等,2015)	202 599幅	–	40	×	×
	CUB-200-2011(Wah等,2011)	11 788幅	–	312	×	×
视频	Bouncing balls(Chang等,2017)	–	32×32	3	√	×
	Sprites(Li和Mandt,2018)	–	64×64	5	√	√
	MMNIST(Srivastava等,2015)	10 000段	64×64	4	√	×
	MUG(Aifanti等,2010)	3 528段	–	2	×	×
	UCF-101(Soomro等,2012)	13 320段	320×240	–	×	×
	Weizmann action(Blank等,2005)	90段	180×144	2	×	×
	KTH(Schuldt等,2004)	–	160×120	3	×	×

注:“–”代表数据缺省,“√”代表该数据集是合成数据集或解耦测评基准数据集;×代表该数据集非合成数据集或解耦测评基准数据集。

的取值。人物角色有9种动作和3个不同的视角。每个视频序列包含8帧,每帧是64×64像素的RGB图像。

3)MMNIST数据集。MMNIST(moving MNIST)(Srivasta等,2015)包含10 000个视频序列,每个序列由20帧组成,每帧分辨率为64×64像素。在每个视频序列中,两个数字独立移动,数字经常相互交叉并从边缘反弹。

4)MUG数据集。MUG(MUG facial expression)(Aifant等,2010)包含3 528段由52名演员表演愤怒、生气、厌恶、开心、伤心和惊讶6种不同面部表情的视频序列。每个视频序列包含50～160帧不等。

5)UCF-101数据集。UCF-101(101 human action classes from videos in the wild)(Soomro等,2012)是从YouTube收集的真实动作视频的动作识别数据集,具有101个动作类别的13 320个视频,并且在相机运动、物体外观和姿势、物体尺度、视点、杂乱背景和照明条件等方面存在较大变化。

6)Weizmann action数据集。Weizmann action(Blank等,2005)包含90段由9个人表演的10种动作的视频序列。10个动作有弯腰、跳千斤顶、向前跳、原地跳、奔跑、侧身奔跑、跳跃、行走,挥动一只手和挥动两只手。视频的分辨率是180×144像素。

7)KTH数据集。KTH(KTH action dataset)(Schuldt等,2004)包含600个由25个人在4个场景表演的6种动作的视频。6种动作分别是拳击、拍手、挥手、慢跑、跑步和步行。视频的速率是25幅/s,分辨率是160×120像素。

4.2　解耦表示的评估

如图6所示,潜在表征的性能评估围绕模块化、紧凑性和明确性3个方面展开。因此,本文按评估内容将常用的解耦度量指标分为模块化指标、紧凑

性指标、明确性指标和综合性指标。其中，综合性指标可以度量超过一种的解耦表征性能。

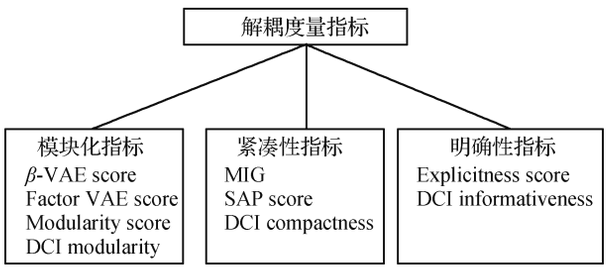


图6　常用的解耦度量指标

Fig. 6　Summary of commonly used DRL metrics

4.2.1　模块化指标

模块化指标主要评估解耦表示的模块化属性，即衡量解耦表示将观测数据的潜在变化因素分离程度。如果潜在表示达到理想的模块化属性，则每个变量（或维度）最多捕捉一个变化因素。常用的模块化解耦指标总结如下：

1）BetaVAE score（z-diff）（Higgins 等，2017）。对于所有生成因素 $\boldsymbol{G}=\{g_1,g_2,\cdots,g_m\}$，每个批次随机固定一个生成因素 g_l 的值，其他生成因素随机取值，生成 N 个数据对 $\{\boldsymbol{x}_1^l,\boldsymbol{x}_2^l\}$。计算每个数据对的潜码的绝对值差异 $Z_{\text{diff}}^l(i)=\left|z_1^l(i)-z_2^l(i)\right|$。如果潜码具有良好的模块化属性，则 Z_{diff}^l 在与固定的生成因素相关的那个维度上应该为零，至少要小于潜码在其他维度上的差异。线性回归器根据每一批里潜码的平均差值 $z_{\text{diff}}^{\mu}=\frac{1}{N}\sum_{i=1}^{N}z_{\text{diff}}^l(i)$ 预测哪个生成因素被固定，预测准确率为指标得分。得分越高说明模块化程度越高。

2）FactorVAE score（Kim 和 Mnih，2018）。FactorVAE score 的计算步骤和思想非常类似于 β-VAE，认为如果观测数据的某个生成因素值相同，则潜码在该因素的对应维度上的编码应该相等。FactorVAE（Kim 和 Mnih，2018）指标与 β-VAE（Higgins 等，2017）指标的区别在于，当一个生成因子被固定时，随机变换其他生成因素，生成单个数据而非成对数据。另一个区别在于 FactorVAE 使用多投票分类器，取方差最小的潜码维度预测固定的生成因素。如果潜在表示完美解耦，则潜码在对应于固定因素的维度上的经验方差应该是 0。

3）Modularity score（Ridgeway 和 Mozer，2018）。首先，计算潜码各维度 $\boldsymbol{z}_i$ 与每个生成因素 $\boldsymbol{v}_j$ 之间的互信息 m_{ij}。理想情况下，如果潜码维度 i 满足模块化属性，它只与一个生成因素有较高的互信息 $\max_j(m_{ij})$，而与其他生成因素的互信息值为零。然后，计算实际互信息值与理想情况下的互信息的偏差 δ_i，具体为

$$\delta_i=\frac{\sum_j(m_{ij}-t_{ij})^2}{\theta_i^2(N-1)},\quad t_{ij}=\begin{cases}\theta_i & v_j=\max_j\left(m_{ij}\right)\\ 0 & \text{其他}\end{cases}\tag{33}$$

式中，m_{ij} 是实际情况的互信息，t_{ij} 是理想情况的互信息，$\theta_i=\max_j(m_{ij})$，N 是生成因素的个数。如果 δ_i 取值为 0，说明潜码维度 i 实现了完美的模块化；如果 δ_i 取值为 1，说明该维度和每个生成因素的互信息相等，同时捕获了多个生成因素的信息。所以使用 $1-\delta_i$ 在潜码各维度上的均值作为整个潜码的模块化分数。

4.2.2　紧凑性指标

紧凑性指标衡量每个潜在变化因素的信息被潜码的单个维度捕获的程度。

1）MIG（mutual information gap）（Chen 等，2019）。计算每个潜码维度 $\boldsymbol{z}_j$ 与生成因素 $\boldsymbol{v}_i$ 之间的互信息 $I\left(\boldsymbol{v}_i,\boldsymbol{z}_j\right)$。对于每个生成因素 $\boldsymbol{v}_i$，潜码各维度与其互信息的最大值记为 $I\left(\boldsymbol{v}_i,\boldsymbol{z}_{\text{最}}\right)$，第二大值记为 $I\left(\boldsymbol{v}_i,\boldsymbol{z}_{\text{次}}\right)$，两者之间的差距经过归一化后即为

$$MIG=\frac{I\left(\boldsymbol{v}_i,\boldsymbol{z}_{\text{最}}\right)-I\left(\boldsymbol{v}_i,\boldsymbol{z}_{\text{次}}\right)}{\sum_{j=1}^{d}I\left(\boldsymbol{v}_i,\boldsymbol{z}_j\right)}\tag{34}$$

所有因素的 MIG 得分的平均值即为最终得分。如果两个最高值十分接近，则说明不只一个潜码维度编码了同一个生成因素，所以潜码具有较低的紧凑性。反之，如果两个最高值差距较大，则说明一个潜码维度足以编码一个变化因素的完整信息，潜码每个维度的编码不同的变化因素，紧凑性高且模块化程度高。

2）SAP score（attribute predictability score）（Kumar 等，2018）。回归器或分类器根据潜码的维度分量 $\boldsymbol{z}_j$ 预测各生成因素 $\boldsymbol{v}_i$ 的真实值，计算一个得分 S_{ij}。若生成因素是一个连续型变量，则 S_{ij} 是线性回归器的回归曲线 R^2 值；若生成因素是一个离散型变量，则 S_{ij} 是决策树平衡分类准确率。最终的 SAP 得分对于每个生成因素，计算两个最高的 S_{ij} 之差，最后求取均值。具体为

$$SAP = \frac{1}{M}\sum_{i}^{M}(S_{i\hat{a}} - S_{i^{\circ}}) \tag{35}$$

式中，$S_{i\hat{a}}$是各潜变量预测生成因素$\boldsymbol{v}_i$的最高得分，$S_{i^{\circ}}$是第2高得分，这与MIG的差距思想相似。

4.2.3 明确性指标

常用的明确性指标为Explicitness score（Ridgeway和Mozer，2018）。

如果潜在表示具有明确性，则从潜码中应该可以恢复生成因素的值。假设生成因素是离散型变量，提出训练一个简单的逻辑回归分类器根据推论得到潜码预测生成因素的真实值。分类器的性能用ROC曲线下面积（ROC area-under-the curve，ROCAUC）衡量。最终的得分为ROCAUC在所有生成因素的所有类别上的平均得分，且得分归一化至[0，1]区间，所以最小取值为0.5。

4.2.4 综合性指标

常用的综合性指标为DCI（disentanglement，compactness，informativeness）。

Eastwood和Williams（2018）提出一个完整的框架评估潜在表示的模块化、紧凑性和明确性3个属性。3个属性指标均训练线性拉索回归器（linear lasso regressor）或随机森林（random forest）用潜码预测生成因素。对于单个生成因素$\boldsymbol{v}_j$，潜码中各维度$\boldsymbol{z}_i$与其的相关程度不同，所以用分类器的权重参数$\boldsymbol{W}$定义一个相对重要性矩阵$\boldsymbol{R}$为$R_{ij} = \left|W_{ij}\right|$，其中$R_{ij}$表示潜码维度$\boldsymbol{z}_i$在预测$\boldsymbol{v}_j$时的相对重要性权重。潜码中的一个维度$\boldsymbol{z}_i$的模块化量化为

$$D_i = 1 - H_K(P_i) \tag{36}$$

式中，$H_K(P_{i\cdot}) = -\sum_{k=0}^{K-1}P_{ik}\log_K P_{ik}$代表熵值，$P_{ij} = R_{ij}/\sum_{k}^{K}R_{ik}$表示潜码维度$\boldsymbol{z}_i$对于预测生成因素$\boldsymbol{v}_j$很重要的“概率”。如果潜码维度$\boldsymbol{z}_i$只对于预测单个生成因子很重要，得分为1；如果$\boldsymbol{z}_i$对于预测所有生成因素同样重要，则得分为0。

潜码中的维度$\boldsymbol{z}_i$关于表达生成因素$\boldsymbol{v}_j$的紧凑指标得分量化为$\boldsymbol{C}_j$，即

$$C_j = 1 - H_D(P_{\cdot j}) \tag{37}$$

式中，$H_D(\tilde{P}_{\cdot j}) = -\sum_{d=0}^{D-1}\tilde{P}_{dj}\log_D \tilde{P}_{ij}$表示$\tilde{P}_{\cdot j}$分布的熵。如果潜码的某单个维度$\boldsymbol{z}_i$完整捕获了生成因素$\boldsymbol{v}_j$的信息，可以预测出$\boldsymbol{v}_j$的真实值，则$C_j$的得分为1。如果需要潜码的所有维度才能预测出$\boldsymbol{v}_j$的真实值，且每个维度对$\boldsymbol{v}_j$预测的贡献相等，则$C_j$为0。

整个潜码$\boldsymbol{z}$对于生成因素$\boldsymbol{v}_j$的明确性得分I由预测误差量化，具体为

$$I = E(\boldsymbol{v}_j, \hat{\boldsymbol{v}}_j) \tag{38}$$

式中，$\boldsymbol{v}_j$是生成因素真实值，$\hat{\boldsymbol{v}}_j$代表预测值，E是一个合适的误差函数。如果模型可以在潜码$\boldsymbol{z}$中明确表示有关生成因素$\boldsymbol{v}_j$的信息，则预测误差应该趋近于0。反之，如果明确性不高，则预测误差应该较高。

4.3 解耦度量指标对比

4.3.1 解耦度量指标的相关一致性对比

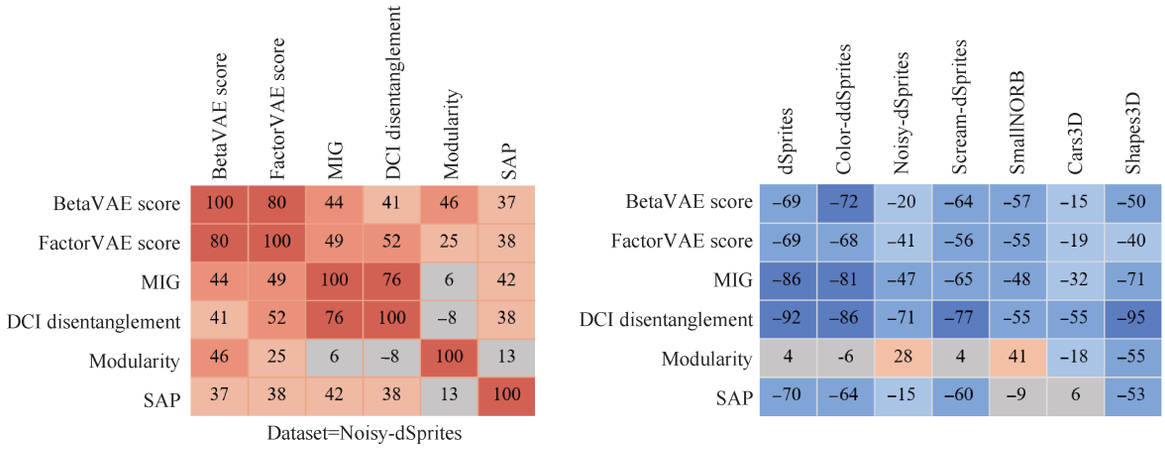
因为对于解耦表示学习没有一个统一化的定义，同时解耦评估指标也没有一个规范化的衡量标准，Locatello等人（2019b）对4.2节介绍的几种常用解耦指标的一致性进行了实验。图7（a）为不同解耦指标在Noisy-dSprites数据集上的斯皮尔曼等级相关（Spearman rank correlation）。结果表明，除了Moularity指标，其他指标在dSprites数据集上均呈现较强相关性，说明虽然大多数解耦指标定义有各种差异，但衡量解耦表征的内容和角度在一定程度上具有一致性。其中，BetaVAE score和FactorVAE score（相关性=80），MIG和DCI disentanglement（相关性=76）这两对指标的等级相关指数较高，说明BetaVAE score和FactorVAE score指标捕捉到的解耦表征概念特别相似。MIG和DCI disentanglement同理。

4.3.2 解耦度量指标的公平性对比

潜在表示的公平性是指该表征是关于观测数据的一种通用表征，可以解决许多与真实生成因素相关的下游任务，而非受限于某一种或几种特定的下游任务。为了探究解耦是否可以在一定程度上减少各潜在表征间的不公平性，Locatello等人（2019a）在多个数据集上对解耦度量指标得分和公平性得分进行了实验。其中，潜在表征的不公平性是通过让训练好的各潜变量模型根据学习到的潜在表征预测出生成因素的值和真实标签的误差来衡量的，具体为

$$unfairness(\hat{y}) = \frac{1}{|S|}\sum_{s}TV(p(\hat{\boldsymbol{y}}), p(\hat{\boldsymbol{y}}\mid \boldsymbol{s}=s)), \forall y \tag{39}$$

式中，S是模型未知的潜在生成因素，即实验者已知的真实标签值，$\hat{y}$是预测出的生成因素的值，TV是总变分（total variation）。该指标得分越低，说明潜变量



(a) 不同解耦度量指标间的等级相关性　　(b) 解耦度量指标和公平性间的等级相关性

图 7　不同的解度量指标在基准数据集上的等级相关性

Fig. 7　Rank correlation of different metrics of DRL on benchmarks((a)the rank correlation between different disentanglement metrics;(b)the rank correlation between disentanglement metrics and fairness)

模型学习到的潜在表示公平性越高。

图 7(b)显示了各评估指标的解耦得分和公平性得分的斯皮尔曼等级相关指数。结果显示，除了 Modularity 指标之外，其他度量指标的解耦得分都与较低的不公平性相关具有一致相关性，说明解耦表示学习在一定程度上可以保证学习到的潜在表征更具有公平性。

5　解耦表示学习的应用

学习高维感官数据的良好表示对于人工智能任务具有根本性的意义。由于解耦表征对观测数据中显著的变化因素进行独立编码，可以更好地解释和控制潜在的变化因素，所以近年来解耦表示常用于解决许多现实世界中的下游任务。本节将总结应用解耦表示学习的常见经典深度学习任务，并分析解耦表示在各深度学习任务中的作用。

5.1　识别分类

解耦表示学习常应用于识别分类相关的深度学习任务，如行人重识别、动作识别(Liu 等，2020；Yang 等，2019；Zhao 等，2018)、人脸识别(Liu 等，2018b；Peng 等，2017；Tran 等，2017；Zhao 等，2019)、步态识别(Li 等，2020e；Zhang 等，2019b)和分类(Jung 等，2020；Ojha 等，2020)。当应用于分类识别问题时，解耦表示模型(Li 等，2020e；Ojha 等，2020；Zhang 等，2019a；Zou 等，2020)借助身份或类别标签作为监督信号，在弱监督或半监督学习方式下，分离出数据本身具有不变性的内在因素(即观测对象的身份特征或类别信息)和变化性的外在因素(如光照、环境、姿势和类别无关信息等)。表 6 展示了不同的行人重识别算法在 Market-1501(Zheng 等，2015)、CUHK03(Chinese University of Hong Kong 03 dataset)(Li 等，2014)、DukeMTMC-reID(Zheng 等，2017)和 MSMT17(multi-scene multitime person ReID dataset)(Wei 等，2018)等 4 个数据集上的表现。其中，R1 是 Rank-1 的简写，表示搜索结果中置信度最高的第 1 幅图识别准确的概率，mAP(mean average precision)表示模型在所有类别上的平均识别准确率。两项指标都是数值越大，算法性能越好。实验结果显示，运用解耦表征学习的算法在 4 个数据集上的总体表现更优，证明了解耦学习表征由于丢弃了具有干扰性的外在因素信息、提炼出类别相关或身份相关的内在因素，所以在一定程度上提高了样本的学习效率，同时对干扰噪声具有更强的鲁棒性。

5.2　数据生成与处理

解耦表示学习也常用于图像生成(Deng 等，2020；Li 等，2020b；Singh 等，2019；Xiao 等，2019a；Yang 等，2019；Yin 等，2019；Zhang 等，2019a)、视频生成(Tulyakov 等，2018；Wang 等，2020b；Zhu 等，2020a)和图像修复(Gilbert 等，2018；Li 等，2020c；Xiao 等，2019b；Liao 等，2019)等深度学习任务。对于图像生成任务，FineGAN(Singh 等，2019)和 MixN-Match(conditional mix-and-match image generation)(Li 等，2020f)可以同时解耦出图像的多个潜在变化

表6　几种行人重识别方法在常用的行人数据集上的识别结果

Table 6　Comparison of state-of-the-art methods on person re-identification dataset　/%

方法类型	算法	Market-1501		CUHK03				DukeMTMC-reID		MSMT17	
				Detected		Labeled					
		R1	mAP	R1	mAP	R1	mAP	R1	mAP	R1	mAP
非解耦方法	HPM(Fu等,2019)	94.2	82.7	63.9	57.5	–	–	86.6	74.3	–	–
	DG-Net(Zheng等,2019)	94.8	86.0	65.6	66.1	–	–	86.6	74.8	–	–
	ICE(Chen等,2021a)	95.1	86.6	–	–	–	–	88.2	76.5	76.4	50.4
	IICS(Xuan等,2021)	89.5	72.9	–	–	–	–	80.0	64.6	56.4	26.9
解耦方法	Pu等(Pu等,2020)	94.7	86.7	70.5	65.8	72.2	67.8	89.05	78.02	79.7	56.3
	IS-GAN(Eom和Ham,2019)	95.2	87.1	74.1	72.5	**72.3**	**68.8**	90.0	79.5	–	–
	Zang等(Zang等,2021)	96.2	**94.9**	–	–	–	–	**92.9**	**91.0**	**88.4**	**79.1**
	Min等(Ren等,2021)	**97.6**	90.8	**95.7**	**89.1**	–	–	–	–	–	–

注:加粗字体表示各列最优结果,"–"代表数据缺省。

因素(图像背景、对象姿态、对象形状和对象外貌),细粒度可控地生成真实多样化的图像。在视频生成中,解耦表示模型(Tulyakov等,2018;Li和Mandt,2018;Zhu等,2020a)将视频分解为时间不变性的静态因素(即目标身份特征)和时间变化性的动态因素(动作、姿势和位置特征),可以生成相同对象但不同动作或者相同动作但不同对象的视频。对于图像修复和图像去模糊等图像处理任务,解耦表示学习(Li等,2020a;Lu等,2019)从受损图像的潜在表示中解耦出失真信号特征或模糊特征并对其单独操作,以实现图像质量复原。

表7展示了不同的条件图像生成算法在CUB(Wah等,2011)、Stanford-Cars(Krause等,2013)和FS-100(few-shots 100 dataset)(Wu等,2019)测试基准上的指标得分。3个指标分别是度量真实图像和生成图像的相似度的FID(frechet inception distance score)、评估生成图像质量的IS(inception score)和生成过程中指定的类标签被用做真实值来计算识别精度(recognition accuracy,RA)。其中,FID数值越小,算法性能越好;IS和RA数值越大,算法性能越好。实验结果显示,采用解耦方法的算法生成的图像通常更接近真实图像、质量更高且更好地保留了条件生成过程中的特定类别信息。

5.3　域适应和风格迁移

不少研究工作将解耦表示学习与域适应(Baktashmotlagh等,2018;Cai等,2019;Zhu等,

表7　几种条件图像生成方法在常用的数据集上的质量对比

Table 7　Comparison of state-of-the-art methods on conditional image generation dataset

方法类型	算法	CUB			Stanford Cars			FS-100		
		FID	IS	RA	FID	IS	RA	FID	IS	RA
非解耦方法	SN-GAN(Miyato等,2018)	160.09	4.21±0.05	–	53.20	2.80±0.05	–	41.26	1.66±0.05	–
	EnhancedTGAN(Wu等,2019)	133.57	4.17±0.03	9.16	105.20	2.43±0.05	3.48	57.58	1.57±0.02	62.69
	R^3-CGAN(Liu等,2020)	88.62	4.43±0.06	8.60	44.57	3.05±0.04	5.48	25.28	1.73±0.02	74.30
解耦方法	FineGAN(Singh等,2019)	46.68	4.62±0.03	–	45.72	2.85±0.04	–	24.63	1.76±0.02	–
	MixNMatch(Li等,2020f)	45.59	**4.78±0.08**	–	45.94	2.60±0.05	–	25.63	1.71±0.05	–
	SSC-GAN(Chen等,2021b)	**20.03**	4.68±0.04	**97.85**	**39.02**	**3.10±0.03**	**87.45**	**20.65**	**1.82±0.03**	**96.86**

注:加粗字体表示各列最优结果。"–"代表数据缺省。

2017)、风格迁移和图像到图像翻译(Gonzalez-Garcia等,2018;Lee等,2018;Liu等,2018d)任务相结合。如Cai等人(2019)和Baktashmotlagh等人(2018)将源域和目标域数据的潜在表示分解为共享变量(语义潜变量)和特有变量(域潜变量)。共享变量代表了源域和目标域之间的公共部分,作为域适应中的可迁移的特征。而特有变量则是每个域的独特风格的特征。模型通过从源域数据中提取域不变性的语义潜变量和目标域数据中的域潜变量特征组合重建,实现源域数据到目标域数据的迁移。解耦表示学习在图像到图像翻译(Gonzalez-Garcia等,2018;Lee等,2018)及风格迁移学习(Kotovenko等,2019;Liu等,2018d)的解决思路与特征和另一图像的风格特征组合生成翻译后的图像或者风格迁移后的图像。表8展示了图像到图像翻译算法在INIT(instance-aware image-to-image translation approach)数据集(Shen等,2019)和Yosemite(summer & winter images dataset at Yosemite)数据集(Zhu等,2017)上的实验结果。其中,INIT数据集选取了sunny→night,sunny→rainy,sunny→cloudy这3个方向,测评指标选取IS、CIS(conditional inception score)和感知损失(learned perceptual image patch similarity,LPIPS)。Yosemite数据集选取了summer→winter,winter→summer两个方向,指标选取FID和LPIPS。IS,CIS,LPIPS数值越大,算法性能越好;FID数值越小,算法性能越好。结果显示,学习解耦表征的算法在整体上的指标得分更优。表明翻译合成的图像更加真实、质量更高、更具有多样性。域适应任务类似,将图像的潜在表示解耦为共享的内容特征和特有的风格特征,通过将某一幅图像的内容特征和另一幅图像的风格特征组合生成翻译后的图像或者风格迁移后的图像。

表8　几种图像到图像翻译方法在常用的数据集上的质量对比

Table 8　Comparison of state-of-the-art methods on image-to-image translation dataset

方法类型	算法	sunny→night			sunny→rainy			sunny→cloudy			summer→winter		winter→summer	
		IS	CIS	LPIPS	IS	CIS	LPIPS	IS	CIS	LPIPS	FID	LPIPS	FID	LPIPS
非解耦方法	CycleGAN(Zhu等,2017)	1.026	0.014	0.016	1.073	0.011	0.008	1.097	0.014	0.011	–	–	–	–
	UNIT(Zhu等,2017)	1.030	0.082	0.067	1.075	0.097	0.062	1.134	0.081	0.068	–	–	–	–
	StarGAN(Choi等,2018)	–	–	–	–	–	–	–	–	–	152.11	0.012	153.79	0.013
	SingleGAN(Yu等,2018)	–	–	–	–	–	–	–	–	–	63.77	0.184	**42.24**	0.188
解耦方法	MUNIT(Huang等,2018)	**1.278**	**1.159**	0.292	1.146	1.012	0.239	1.095	1.008	0.211	84.43	0.166	73.82	0.134
	DRIT(Lee等,2018)	1.224	1.058	0.231	**1.207**	1.007	0.173	1.104	**1.025**	0.166	58.70	0.205	53.79	0.192
	INIT(Shen等,2019)	1.118	1.060	**0.330**	1.152	**1.036**	**0.267**	**1.460**	1.016	**0.224**	–	–	–	–
	DMIT(Yu等,2019)	–	–	–	–	–	–	–	–	–	**58.46**	**0.302**	48.02	**0.292**

注:加粗字体表示各列最优结果。"–"代表数据缺省。

5.4　跨模态信息检索

跨模态检索任务是给定来自源模态的查询条目,需要在目标模态中检索出与其相关的条目。传统跨模态检索方法将所有模态的数据映射到一个公共空间,跨模态数据的表示是高度耦合的。然而不同模态的数据具有不同的属性,即存在异构差距,因此它们之间很难直接建立联系。一些解耦表示学习工作,如IIAE(Hwang等,2020)、PDFD(progressive domainindependent feature decomposition)(Xu等,2021)和Guo等人(2019)认为不同模式下存在一部分在共享的信息,并学习一种解耦表示将跨模态数据信息分解为模态共享信息和模态特有信息。其中,模态共享信息通常为语义特征,是建立跨模式连接的基础,故被投影到一个公共空间。模态特有信息则包含每个模态特有的变化因素,对于跨模态检索不利。表9展示了不同的零样本草图检索方法在Sketchy(Sangkloy等,2016)和TUBerlin(Eitz等,2012)数据集上的平均检索精度mAP和置信度前100个的正样本概率Prec@100。两项指标都是数值越大,算法性能越好。结果显示,应用解耦表示的检

表9　零样本草图检索方法在常用的数据集上的质量对比

Table 9　Comparison of state-of-the-art methods on zero-shot sketch-based image retrieval dataset

方法类型	算法	特征维度	Sketchy Ext.		TUBerlin Ext.	
			mAP	Prec@100	mAP	Prec@100
非解耦方法	SEM-PCYC(Dutta等,2019)	64	0.349	0.463	0.297	0.426
	SketchGCN(Zhang等,2020)	2 048	0.382	0.538	0.324	0.505
	OCEAN(Zhu等,2020a)	64	0.462	0.590	0.333	0.476
	NAVE(Wang等,2021)	64	0.508	0.632	0.412	0.519
		512	0.613	0.725	0.493	0.607
解耦方法	StyleGuide(Dutta等,2021)	300	0.376	0.484	0.254	0.355
	IIAE(Hwang等,2020)	64	0.573	0.695	-	-
	Deng等(Deng等,2020)	354	0.523	0.616	0.424	0.517
	TCN(Wang等,2021)	64	0.488	0.644	0.381	0.506
		512	**0.616**	**0.763**	**0.495**	**0.616**

注:加粗字体表示各列最优结果。"-"代表数据缺省。

索方法在两个数据集上的平均指标得分更高。证明解耦表示学习利用解耦所得的域共享信息检索,可以有效抑制域特有属性因素的干扰,从而提高检索算法的鲁棒性和语义检索准确率。

5.5　医学图像分割

医学图像分割是指根据医学图像的某种相似性特征(如亮度、颜色、纹理、面积、形状、位置、局部统计特征或频谱特征等)将医学图像划分为若干个互不相交的"连通"的区域的过程。医学图像具有较高的复杂性且缺少简单的线性特征,且分割准确率受到部分容积效应、灰度不均匀性、伪影和不同软组织间灰度的接近性等因素的影响,因此从医学图像中自动分割出目标是一个具有挑战性的问题。核磁共振影像(magnetic resonance image, MRI)较计算机断层扫描(computed tomography, CT)影像包含更多复杂的数据信息,但却更难获得。Chartsias等人(2019, 2020)和Yang等人(2019)利用多模态的心脏影像和肝脏影像,通过解耦表示算法来学习一种跨模态医学分割模型。在Chartsias等人(2019)的模型中,空间内容因素被表示为多类别语义图,与输入图像具有像素级的对应关联,将其输入进一个分割网络中,可产生多类别分别掩膜,并且可适用于表示任何模态下的解剖语义。表10展示了几种医学图像分割方法在CT-MR(Roberson等,2005)数据集上对左心室腔(left ventricular cavity, LV)和左心室心肌(left ventricular myocardium, MYO)进行CT→MR和MR→CT两个方向跨域分割的实验结果。为了评估分割准确性,本文选取用于衡量预测和黄金分割之间的重叠的骰子系数(Dice, DS)和用于评估模型在边界的表现的平均对称面距离(average symmetric surface distance, ASD)两个指标。DS数值越大,算法性能越好;ASD数值越小,算法性能越好。整体实验结果表明,解耦表示学习在两个跨域方向上的分割准确度较非解耦算法更高。解耦表征学习将多模态2维医学图像被分解为空间内容因素和非空间风格因素。空间内容因素可视为多个模态间共享的内容特征,保留了解剖信息,可以提高图像分割的样本效率。

6　解耦学习未来展望与总结

6.1　未来研究方向

1)提出通用化的、公平性更高的解耦表示学习算法。虽然现有许多研究工作提出了性能良好的解耦表示算法,然而大多数的解耦表示模型局限性较强,只适用于特定数据或特定下游任务,且解耦模型的性能在很大程度上受随机种子和超参数设置的影响,存在训练不稳定的问题(Locatello等,2019b)。当模型迁移应用至其他类似的场景或数据集,超参数和随机种子发生变化时,算法的性能常会出现一

表 10　医学图像分割方法在常用的数据集上的质量对比

Table 10　Comparison of state-of-the-art methods on cardiac image segmentation dataset

方法类型	算法	CT→MR				MR→CT			
		LV		MYO		LV		MYO	
		DS/%	ASD/mm	DS/%	ASD/mm	DS/%	ASD/mm	DS/%	ASD/mm
非解耦方法	CycleSe(Zhu 等,2017)	81.3±11.8	6.6±3.6	53.2±11.7	11.8±5.1	79.3±15.3	8.3±3.9	51.3±15.4	6.6±3.8
	PnP-AdaNet(Dou 等,2019)	86.2±6.46	**2.74±1.04**	57.9±8.43	**2.46±0.611**	78.31±8.4	3.88±4.09	62.8±8.24	3.091±59
	Ouyang 等(Ouyang 等,2022)	–	–	–	–	61.1±26.9	9.8±5.3	75.8±8.7	5.6±1.2
	SIFA(Chen 等,2019)	87.6±8.9	4.6±2.3	67.3±11.4	8.2±5.3	82.6±12.6	7.8±3.0	56.6±12.4	6.8±3.8
解耦方法	MMRegNet(Ding 等,2021)	–	–	–	–	80.3±7.2	**3.46±1.30**	62.9±8.62	**3.01±0.74**
	CFDnet(Wu 和 Lu,2020)	**88.7±10.6**	2.99±2.79	67.9±8.62	3.40±2.75	81.9±18.2	3.64±3.94	62.9±10.9	3.16±1.18
	DDA-GAN(Chen 等,2021c)	–	–	–	–	78.5±6.9	5.4±1.4	**77.8±10.2**	5.2±1.9
	DDFSeg(Pei 等,2021)	87.7±10.4	3.8±1.9	**71.3±10.1**	9.7±5.7	**83.5±16.0**	8.3±4.2	66.9±11.0	6.8±4.6

注:加粗字体表示各列最优结果。“–”代表数据缺省。

定程度的退化。因此,实现适用范围更广、泛化能力更强、迁移和可扩展性更高、对超参数和随机种子设定更加鲁棒的算法模型是解耦表示学习面临的重要挑战之一。

2)继续深入探究解耦表示学习背后的机理。研究者通常认为解耦化的表示对于解决许多现实世界中的下游任务具有帮助。近年来,许多研究工作运用解耦表示学习来解决各类现实世界的人工智能问题,并且都取得了可观性的突破。然而,并非所有的解耦表示学习工作都能通过严格的理论,证明所提出的潜变量模型可以正确恢复出潜在变化因素的真实分布。部分生成领域的解耦表示学习工作能通过实验数据证明所提出的潜变量模型在一定程度上提高算法性能,但其背后的原理却未能在论文中进行严格地数学推导和证明。因此,深入分析探究解耦表示学习背后的数学原理是其未来发展的一个重要方向。

3)构建质量更高、更贴近自然数据的解耦测评基准数据集。由于合成数据集的成本低、易生成且独立的变化因素易控制,解耦表示学习基准数据集基本上均为合成数据集。然而当前基准数据集如 dSprites(Matthey 等,2017),smallNORB(LeCun 等,2004),3D Shapes(Matthey 等,2017)的图像分辨率较小,生成因素个数较少,且模拟环境下的合成图像与真实世界图像相比真实性差异较大。只有很少的解耦表示工作构建了场景复杂度更高、更具有照片真实性、变化因素更丰富的高分辨率解耦基准数据集,如 Falcor3D(Gondal 等,2019),Issac3D(Gondal 等,2019)和 MPI3D(Nie 等,2020a)。然而这些新提出的解耦基准数据集还未受到广泛关注和应用。在今后的解耦表示发展中,缩小合成数据集与现实数据集之间的差距,提供具有更复杂的形状和纹理的 3 维重建或场景渲染的测评基准数据集是研究者亟待解决的问题之一。

4)规范解耦表示的性能评估。虽然当前有许多工作提出了创新性的解耦度量指标,但当前还未形成一套广泛认可的系统化、规范化的解耦度量框架和度量标准(Carbonneau 等,2022)。部分解耦评估指标给出了量化评估的具体步骤,但未能明确其度量潜在表征的何种性能,也未能明确进行评估时需满足的实验条件。因此,在很多解耦表示模型在对比实验中的度量有失公平。形成规范化、统一化的解耦性能评估标准和框架有利于解耦表示学习更好地发展,是解耦表示学习研究中的一项重要内容。

6.2　总结

本文首先介绍了解耦表示学习的概念以及当前的研究现状,并归纳分析了解耦表示的因果机制及其 3 个标准属性。然后将当前的解耦表示学习工作归纳为维度解耦、语义解耦、层级解耦和非线性独立成分分析 4 类,分别从概率模型、特点和适用范围等方面进行分析和对比。根据解耦表示模块化、紧凑性和明确性 3 个属性,对损失函数和评估指标进行

分类梳理,对比分析各损失函数的适用范围和局限性,并对几种常用的评估指标进行了深入地探讨。最后,从算法泛化性、原理探究、测评基准数据集质量和度量指标的规范性4个方面分析了解耦表示学习面临的挑战和未来的发展趋势。可以看出,解耦表示学习具有重要的实际应用价值和巨大的发展空间。

参考文献(References)

Achille A, Eccles T, Matthey L, Burgess C, Watters N, Lerchner A and Higgins I. 2018. Life-long disentangled representation learning with cross-domain latent homologies//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montreal, Canada: Curran Associates Inc

Aifanti N, Papachristou C and Delopoulos A. 2010. The MUG facial expression database//Proceedings of the 11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10. Desenzano del Garda, Italy: IEEE: 1-4

Aubry M, Maturana D, Efros A A, Russell B C and Sivic J. 2014. Seeing 3D chairs: exemplar part-based 2D-3D alignment using a large dataset of CAD models//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE: 3762-3769 [DOI: 10.1109/CVPR.2014.487]

Bai J W, Kong S F and Gomes C. 2020a. Disentangled variational auto-encoder based multi-label classification with covariance-aware multivariate probit model//Proceedings of the 29th International Joint Conference on Artificial Intelligence. Yokohama, Japan: IJCAI. org: 4313-4321 [DOI: 10.24963/ijcai.2020/595]

Bai Y, Lou Y H, Dai Y X, Liu J, Chen Z Q and Duan L Y. 2020b. Disentangled feature learning network for vehicle re-identification//Proceedings of the 29th International Joint Conference on Artificial Intelligence. Yokohama, Japan: IJCAI. org: 474-480 [DOI: 10.24963/ijcai.2020/66]

Baktashmotlagh M, Faraki M, Drummond T and Salzmann M. 2018. Learning factorized representations for open-set domain adaptation. [EB/OL]. [2022-01-21]. https://arxiv.org/pdf/1805.12277.pdf

Bass C, da Silva M, Sudre C, Tudosiu P D, Smith S M and Robinson E C. 2020. ICAM: interpretable classification via disentangled representations and feature attribution mapping//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 7697-7709

Bengio Y, Courville A and Vincent P. 2013. Representation learning: a review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(8): 1798-1828 [DOI: 10.1109/TPAMI.2013.50]

Bepler T, Zhong E D, Kelley K, Brignole E and Berger B. 2019. Explicitly disentangling image content from translation and rotation with spatial-VAE//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Curran Associates, Inc.: 15435-15445

Bi S, Sunkavalli K, Perazzi F, Shechtman E, Kim V G and Ramamoorthi R. 2019. Deep CG2Real: synthetic-to-real translation via image disentanglement//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 2730-2739 [DOI: 10.1109/ICCV.2019.00282]

Blank M, Gorelick L, Shechtman E, Irani M and Basri R. 2005. Actions as space-time shapes//Proceedings of the 10th IEEE International Conference on Computer Vision. Beijing, China: IEEE: 1395-1402 [DOI: 10.1109/ICCV.2005.28]

Bouchacourt D, Tomioka R and Nowozin S. 2018. Multi-level variational autoencoder: learning disentangled representations from grouped observations. Proceedings of 2018 AAAI Conference on Artificial Intelligence, 32 (1): 2095-2102 [DOI: 10.1609/aaai. v32i1.11867]

Bromley J, Guyon I, LeCun Y, Säckinger E and Shah R. 1993. Signature verification using a "Siamese" time delay neural network//Proceedings of the 6th International Conference on Neural Information Processing Systems. Denver, Colorado, USA: Morgan Kaufmann Publishers Inc.: 737-744

Burgess C and Kim H. 2018. 3D shapes dataset [EB/OL]. [2022-01-21]. https://github.com/deepmind/3d-shapes

Burgess C P, Higgins I, Pal A, Matthey L, Watters N, Desjardins G and Lerchner A. 2018. Understanding disentangling in β-VAE [EB/OL]. [2022-01-21]. https://arxiv.org/pdf/1804.03599.pdf

Cai R C, Li Z J, Wei P F, Qiao J, Zhang K and Hao Z F. 2019. Learning disentangled semantic representation for domain adaptation//Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao, China: IJCAI. org: 2060-2066 [DOI: 10.24963/ijcai.2019/285]

Carbonneau M A, Zaïdi J, Boilard J and Gagnon G. 2022. Measuring disentanglement: a review of metrics.IEEE Transactions on Neural Networks and Learning Systems. 2022: 1-15 [DOI: 10.1109/TNNLS.2022.3218982]

Chang M B, Ullman T, Torralba A and Tenenbaum J B. 2017. A compositional object-based approach to learning physical dynamics [EB/OL]. [2022-01-21]. http://arxiv.org/pdf/1612.00341.pdf

Chartsias A, Joyce T, Papanastasiou G, Semple S, Williams M, Newby D E, Dharmakumar R and Tsaftaris S A. 2019. Disentangled representation learning in cardiac image analysis. Medical Image Analysis, 58: #101535 [DOI: 10.1016/j.media.2019.101535]

Chartsias A, Papanastasiou G, Wang C J, Stirrat C, Semple S, Newby D, Dharmakumar R and Tsaftaris S A. 2020. Multimodal cardiac segmentation using disentangled representation learning//Proceedings of the 10th International Workshop on Statistical Atlases and Computational Models of the Heart. Multi-Sequence CMR Segmen-

tation, CRT-EPiggy and LV Full Quantification Challenges. Shenzhen, China: Springer: 128-137 [DOI: 10.1007/978-3-030-39074-7_14]

Chen H Y, Chen F and He H J. 2021a. SSC-GAN: a novel gan based on the same solution constraints of first-order ODEs. International Journal of Pattern Recognition and Artificial Intelligence. 35(11): #2152018 [DOI: 10.1142/S0218001421530062]

Chen H, Lagadec B and Bremond F. 2021b. ICE: inter-instance contrastive encoding for unsupervised person re-identification//Proceedings of 2021 IEEE International Conference on Computer Vision. IEEE: 14960-14969

Chen R T Q, Li X C, Grosse R and Duvenaud D. 2019. Isolating sources of disentanglement in variational autoencoders [EB/OL]. [2022-01-21]. https://arxiv.org/pdf/1802.04942.pdf

Chen X, Duan Y, Houthooft R, Schulman J, Sutskever I and Abbeel P. 2016. InfoGAN: interpretable representation learning by information maximizing generative adversarial nets//Proceedings of the 30th International Conference on Neural Information Processing Systems. Barcelona, Spain: Curran Associates Inc.: 2180-2188

Chen X, Lian C F, Wang L, Deng H N, Kuang T S, Fung S H, Gateno J, Shen D G, Xia J J and Yap P T. 2021c. Diverse data augmentation for learning image segmentation with cross-modality annotations//Medical Image Analysis. 71: #102060 [DOI: 10.1016/j.media.2021.102060]

Choi Y J, Choi M J, Kim M Y, Ha J W, Kim S H and Choo J. 2018. StarGAN: unified generative adversarial networks for multi-domain image-to-image translation//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 8789-8797

Cohen G, Afshar S, Tapson J and van Schaik A. 2017. EMNIST: an extension of MNIST to handwritten letters//Proceedings of 2017 International Joint Conference on Neural Networks (IJCNN). Anchorage, USA: IEEE: 2921-2926 [DOI: 10.1109/IJCNN.2017.7966217]

Creager E, Madras D, Jacobsen J H, Weis M A, Swersky K, Pitassi T and Zemel R. 2019. Flexibly fair representation learning by disentanglement//Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA: PMLR: 1436-1445

Deng Y, Yang J L, Chen D, Wen F and Tong X. 2020. Disentangled and controllable face image generation via 3D imitative-contrastive learning//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 5153-5162 [DOI: 10.1109/CVPR42600.2020.00520]

Denton E and Birodkar V. 2017. Unsupervised learning of disentangled representations from video//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 4417-4426

Detlefsen N S and Hauberg S. 2019. Explicit disentanglement of appearance and perspective in generative models//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Curran Associates, Inc.: 1018-1028

Ding W, Li L, Huang L and Zhuang X. 2022. Unsupervised multi-modality registration network based on spatially encoded gradient information//Statistical Atlases and Computational Models of the Heart. Multi-Disease, Multi-View, and Multi-Center Right Ventricular Segmentation in Cardiac MRI Challenge. Strasbourg, France: Cham: Springer International Publishing: 151-159 [DOI: 10.1007/978-3-030-93722-5_17]

Ding Z, Xu Y F, Xu W J, Parmar G, Yang Y, Welling M and Tu Z W. 2020. Guided variational autoencoder for disentanglement learning//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE: 7917-7926 [DOI: 10.1109/CVPR42600.2020.00794]

Dinh L, Sohl-Dickstein J and Bengio S. 2017. Density estimation using real NVP [EB/OL]. [2022-01-21]. https://arxiv.org/pdf/1605.08803.pdf

Dou Q, Ouyang C, Chen C, Chen H, Glocker B, Zhuang X H and Heng P A. 2019. PnP-AdaNet: plug-and-play adversarial domain adaptation network at unpaired cross-modality cardiac segmentation. IEEE Access. 7: 99065-99076 [DOI: 10.1109/ACCESS.2019.2929258]

Duan B Y, Fu C Y, Li Y, Song X G and He R. 2020. Cross-spectral face hallucination via disentangling independent factors//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 7927-7935 [DOI: 10.1109/CVPR42600.2020.00795]

Dupont E. 2018. Learning disentangled joint continuous and discrete representations//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montreal, Canada: Curran Associates Inc.: 708-718

Dutta A and Akata Z. 2019. Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 5089-5098

Dutta T, Singh A and Biswas S. 2021. StyleGuide: zero-shot sketch-based image retrieval using style-guided image generation. IEEE Transactions on Multimedia. 23: 2833-2842 [DOI: 10.1109/TMM.2020.3017918]

Eastwood C and Williams C K I. 2018. A framework for the quantitative evaluation of disentangled representations//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada: OpenReview.net

Eitz M, Richter R, Boubekeur T and Hildebrand K. 2012. Sketch-based shape retrieval. ACM Transactions on graphics (TOG). 31(4): 1-10 [DOI: 10.1145/2185520.2185527]

Eom C and Ham B. 2019. Learning disentangled representation for robust person re-identification//Proceedings of the 33rd International Conference on Neural Information Processing Systems.Curran

Associates, Inc.: 5297-5308

Esmaeili B, Wu H, Jain S, Bozkurt A, Siddharth N, Paige B, Brooks D H, Dy J and Van de Meent J W. 2019. Structured disentangled representations//Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics. Naha, Japan: PMLR: 2525-2534

Estermann B, Marks M and Yanik M F. 2020. Robust disentanglement of a few factors at a time using rPU-VAE//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 13387-13398

Fidler S, Dickinson S and Urtasun R. 2012. 3D object detection and viewpoint estimation with a deformable 3D cuboid model//Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe, USA: Curran Associates Inc.: 611-619

Fraccaro M, Kamronn S, Paquet U and Winther O. 2017. A disentangled recognition and nonlinear dynamics model for unsupervised learning//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 3604-3613

Fu Y, Wei Y, Zhou Y, Shi H, Huang G, Wang X, Yao Z and Huang T. 2019. Horizontal pyramid matching for person re-identification. Proceedings of 2019 AAAI Conference on Artificial Intelligence, 33(1), 8295-8302 [DOI: 10.1609/aaai.v33i01.33018295]

Gilbert A, Collomosse J, Jin H L and Price B. 2018. Disentangling structure and aesthetics for style-aware image completion//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 1848-1856 [DOI: 10.1109/CVPR.2018.00198]

Gondal M W, Wüthrich M, Miladinović Đ, Locatello F, Breidt M, Volchkov V, Akpo J, Bachem O, Schölkopf B and Bauer S. 2019. On the transfer of inductive bias from simulation to the real world: a new disentanglement dataset//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates, Inc.: 15740-15751

Gonzalez-Garcia A, Van de Weijer J and Bengio Y. 2018. Image-to-image translation for cross-domain disentanglement//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montreal, Canada: Curran Associates Inc.: 1294-1305

Goodfellow I J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A and Bengio Y. 2014. Generative adversarial nets//Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press: 2672-2680

Gowal S, Qin C L, Huang P S, Cemgil T, Dvijotham K, Mann T and Kohli P. 2020. Achieving robustness in the wild via adversarial mixing with disentangled representations//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 1208-1217 [DOI: 10.1109/CVPR42600.2020.00129]

Grathwohl W and Wilson A. 2016. Disentangling space and time in video with hierarchical variational auto-encoders [EB/OL]. [2022-01-21]. https://arxiv.org/pdf/1612.04440.pdf

Gulrajani I, Kumar K, Ahmed F, Taïga A A, Visin F, Vázquez D and Courville A C. 2016. PixelVAE: a latent variable model for natural images. [EB/OL]. [2022-01-21]. https://arxiv.org/pdf/1611.05013.pdf

Guo W K, Huang H B, Kong X W and He R. 2019. Learning disentangled representation for cross-modal retrieval with deep mutual information estimation//Proceedings of the 27th ACM International Conference on Multimedia. Nice, France: ACM: 1712-1720 [DOI: 10.1145/3343031.3351053]

Hadad N, Wolf L and Shahar M. 2018. A two-step disentanglement method//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 772-780 [DOI: 10.1109/CVPR.2018.00087]

Hamaguchi R, Sakurada K and Nakamura R. 2019. Rare event detection using disentangled representation learning//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 9319-9327 [DOI: 10.1109/CVPR.2019.00955]

Higgins I, Amos D, Pfau D, Racaniere S, Matthey L, Rezende D and Lerchner A. 2018. Towards a definition of disentangled representations [EB/OL]. [2022-01-21]. https://arxiv.org/pdf/1812.02230.pdf

Higgins I, Matthey L, Pal A, Burgess C, Glorot X, Botvinick M, Mohamed S and Lerchner A. 2017. β-VAE: learning basic visual concepts with a constrained variational framework//Proceedings of the 5th International Conference on Learning Representations. Toulon, France: OpenReview. net

Hinton G E and Salakhutdinov R R. 2006. Reducing the dimensionality of data with neural networks. Science, 313(5786): 504-507 [DOI: 10.1126/science.1127647]

Hochreiter S and Schmidhuber J. 1997. Long short-term memory. Neural Computation, 9(8): 1735-1780 [DOI: 10.1162/neco.1997.9.8.1735]

Hsieh J T, Liu B B, Huang D A, Li F F and Niebles J C. 2018. Learning to decompose and disentangle representations for video prediction//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montreal, Canada: Curran Associates Inc.: 515-524

Hsu W N, Zhang Y and Glass J. 2017. Unsupervised learning of disentangled and interpretable representations from sequential data//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc.: 1876-1887

Huang X, Liu M Y, Belongie S and Kautz J. 2018. Multimodal unsupervised image-to-image translation//Proceedings of the 15th European

Conference on Computer Vision (ECCV). Munich, Germany: Springer: 172-189

Hwang H, Kim G H, Hong S and Kim K E. 2020. Variational interaction information maximization for cross-domain disentanglement//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 22479-22491

Jiang Z H, Wu Q Y, Chen K Y and Zhang J Y. 2019. Disentangled representation learning for 3D face shape//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 11949-11958 [DOI: 10.1109/CVPR.2019.01223]

Jung D, Lee J, Yi J H and Yoon S. 2020. ICAPS: an interpretable classifier via disentangled capsule networks//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 314-330 [DOI: 10.1007/978-3-030-58529-7_19]

Kim H and Mnih A. 2018. Disentangling by factorising//Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden: PMLR: 2649-2658

Khemakhem I, Kingma D, Monti R and Hyvarinen A. 2020. Variational autoencoders and nonlinear ICA: a unifying framework//Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics. Palermo, Italy: PMLR: 2207-2217

Kingma D P and Welling M. 2013. Auto-encoding variational Bayes [EB/OL]. [2022-01-21]. https://arxiv.org/pdf/1312.6114v1.pdf

Klindt D, Schott L, Sharma Y, Ustyuzhaninov I, Brendel W, Bethge M and Paiton D. 2021. Towards nonlinear disentanglement in natural data with temporal sparse coding. [EB/OL]. [2022-01-21]. https://arxiv.org/pdf/2007.10930.pdf

Kondo R, Kawano K, Koide S and Kutsuna T. 2019. Flow-based image-to-image translation with feature disentanglement//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates, Inc.: 4168-4178

Kotovenko D, Sanakoyeu A, Lang S and Ommer B. 2019. Content and style disentanglement for artistic style transfer//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 4421-4430 [DOI: 10.1109/ICCV.2019.00452]

Krause J, Stark M, Deng J and Li F F. 2013. 3D object representations for fine-grained categorization//Proceedings of 2013 IEEE International Conference on Computer Vision (ICCV) Workshops. Sydney, Australia: IEEE: 554-561

Kulkarni T D, Whitney W F, Kohli P and Tenenbaum J B. 2015. Deep convolutional inverse graphics network//Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press: 2539-2547

Kumar A, Sattigeri P and Balakrishnan A. 2018. Variational inference of disentangled latent concepts from unlabeled observations [EB/OL]. [2022-01-21]. https://arxiv.org/pdf/1711.00848.pdf

Lai C S, You Z Z, Huang C C, Tsai Y H and Chiu W C. 2020. Colorization of depth map via disentanglement//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 450-466 [DOI: 10.1007/978-3-030-58571-6_27]

LeCun Y, Bottou L, Bengio Y and Haffner P. 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11): 2278-2324 [DOI: 10.1109/5.726791]

LeCun Y, Huang F J and Bottou L. 2004. Learning methods for generic object recognition with invariance to pose and lighting//Proceedings of 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington, USA: IEEE: 97-104 [DOI: 10.1109/CVPR.2004.144]

Lee H Y, Tseng H Y, Huang J B, Singh M and Yang M H. 2018. Diverse image-to-image translation via disentangled representations//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany: Springer: 36-52 [DOI: 10.1007/978-3-030-01246-5_3]

Lee W, Kim D, Hong S and Lee H. 2020. High-fidelity synthesis with disentangled representation//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 157-174 [DOI: 10.1007/978-3-030-58574-7_10]

Li P P, Huang H B, Hu Y B, Wu X, He R and Sun Z N. 2020a. Hierarchical face aging through disentangled latent characteristics//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 86-101 [DOI: 10.1007/978-3-030-58580-8_6]

Li P P, Liu Y L, Shi H L, Wu X, Hu Y B, He R and Sun Z N. 2020b. Dual-structure disentangling variational generation for data-limited face parsing//Proceedings of the 28th ACM International Conference on Multimedia. Seattle, USA: ACM: 556-564 [DOI: 10.1145/3394171.3413919]

Li S, Hooi B and Lee G H. 2020c. Identifying through flows for recovering latent representations [EB/OL]. [2022-01-21]. https://arxiv.org/pdf/1909.12555.pdf

Li W, Zhao R, Xiao T and Wang X. 2014. DeepReID: deep filter pairing neural network for person re-identification//Proceedings of 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE: 152-159

Li X, Jin X, Lin J X, Liu S, Wu Y J, Yu T, Zhou W and Chen Z B. 2020d. Learning disentangled feature representation for hybrid-distorted image restoration//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 313-329 [DOI: 10.1007/978-3-030-58526-6_19]

Li X, Makihara Y, Xu C, Yagi Y and Ren M W. 2020e. Gait recognition via semi-supervised disentangled representation learning to identity and covariate features//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 13306-13316 [DOI: 10.1109/CVPR42600.2020.01332]

Li Y H, Singh K K, Ojha U and Lee Y J. 2020f. MixNMatch: multifactor disentanglement and encoding for conditional image generation//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 8036-8045 [DOI: 10.1109/CVPR42600.2020.00806]

Li Y Z and Mandt S. 2018. Disentangled sequential autoencoder//Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden: PMLR: 5670-5679

Li Z Y, Murkute J V, Gyawali P K and Wang L W. 2020g. Progressive learning and disentanglement of hierarchical representations [EB/OL]. [2022-01-21]. https://arxiv.org/pdf/2002.10549.pdf

Liao L, Hu R M, Xiao J and Wang Z Y. 2019. Artist-Net: decorating the inferred content with unified style for image inpainting. IEEE Access, 7: 36921-36933 [DOI: 10.1109/ACCESS.2019.2905268]

Liu A H, Liu Y C, Yeh Y Y and Wang Y C F. 2018a. A unified feature disentangler for multi-domain image translation and manipulation//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montreal, Canada: Curran Associates Inc.: 2595-2604

Liu F, Zhu R H, Zeng D, Zhao Q J and Liu X M. 2018b. Disentangling features in 3D face shapes for joint face reconstruction and recognition//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 5216-5225 [DOI: 10.1109/CVPR.2018.00547]

Liu Y, Wang Z W, Jin H L and Wassell I. 2018c. Multi-task adversarial network for disentangled feature learning//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 3743-3751 [DOI: 10.1109/CVPR.2018.00394]

Liu Y, Wei F Y, Shao J, Sheng L, Yan J J and Wang X G. 2018e. Exploring disentangled feature representation beyond face identification//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 2080-2089 [DOI: 10.1109/CVPR.2018.00222]

Liu Y C, Yeh Y Y, Fu T C, Wang S D, Chiu W C and Wang Y C F. 2018d. Detach and adapt: learning cross-domain disentangled deep representation//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 8867-8876 [DOI: 10.1109/CVPR.2018.00924]

Liu Z W, Luo P, Wang X G and Tang X O. 2015. Deep learning face attributes in the wild//Proceedings of 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE: 3730-3738 [DOI: 10.1109/ICCV.2015.425]

Liu Z Y, Zhang H W, Chen Z H, Wang Z Y and Ouyang W L. 2020. Disentangling and unifying graph convolutions for skeleton-based action recognition//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 140-149 [DOI: 10.1109/CVPR42600.2020.00022]

Locatello F, Abbati G, Rainforth T, Bauer S, Schölkopf B and Bachem O. 2019a. On the fairness of disentangled representations//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates, Inc.: 14611-14624

Locatello F, Bauer S, Lucic M, Raetsch G, Gelly S, Schölkopf B and Bachem O. 2019b. Challenging common assumptions in the unsupervised learning of disentangled representations//Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA: Curran Associates, Inc.: 7247-7283

Lorenz D, Bereska L, Milbich T and Ommer B. 2019. Unsupervised part-based disentangling of object shape and appearance//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 10947-10956 [DOI: 10.1109/CVPR.2019.01121]

Lu B Y, Chen J C and Chellappa R. 2019. Unsupervised domain-specific deblurring via disentangled representations//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 10217-10226 [DOI: 10.1109/CVPR.2019.01047]

Ma J X, Zhou C, Cui P, Yang H X and Zhu W W. 2019. Learning disentangled representations for recommendation//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver Canada: Curran Associates, Inc.: 5711-5722

Ma L Q, Sun Q R, Georgoulis S, Van Gool L, Schiele B and Fritz M. 2018. Disentangled person image generation//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 99-108 [DOI: 10.1109/CVPR.2018.00018]

Massagué A C, Zhang C, Feric Z, Camps O and Yu R. 2020. Learning disentangled representations of video with missing data//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 3625-3635

Matthey L, Higgins I, Hassabis D and Lerchner A. 2017. dSprites: disentanglement testing sprites dataset [EB/OL]. [2022-01-21]. https://github.com/deepmind/dsprites-dataset/

Miyato T, Kataoka T, Koyama M and Yoshida Y. 2018. Spectral normalization for generative adversarial networks// Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden: PMLR

Netzer Y, Wang T, Coates A, Bissacco A, Wu B and Ng A Y. 2011. Reading digits in natural images with unsupervised feature learning//NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011

Nie Q, Liu Z W and Liu Y H. 2020a. Unsupervised 3D human pose representation with viewpoint and pose disentanglement//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 102-118 [DOI: 10.1007/978-3-030-58529-7_7]

Nie W L, Karras T, Garg A, Debnath S, Patney A, Patel A B and

Anandkumar A. 2020b. Semi-supervised StyleGAN for disentanglement learning//Proceedings of the 37th International Conference on Machine Learning. Virtual: JMLR.org: 7360-7369

Niu X S, Yu Z T, Han H, Li X B, Shan S G and Zhao G Y. 2020. Video-based remote physiological measurement via cross-verified feature disentangling//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 295-310 [DOI: 10.1007/978-3-030-58536-5_18]

Ojha U, Singh K K, Hsieh C J and Lee Y J. 2020. Elastic-InfoGAN: unsupervised disentangled representation learning in class-imbalanced data//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 18063-18075

Ouyang C, Biffi C, Chen C, Kart T, Qiu H Q and Rueckert D. 2022. Self-supervised learning for few-shot medical image segmentation. IEEE Transactions on Medical Imaging. 41(7): 1837-1848 [DOI: 10.1109/TMI.2022.3150682]

Painter M, Hare J and Prugel-Bennett A. 2020. Linear disentangled representations and unsupervised action estimation//Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 13297-13307

Paysan P, Knothe R, Amberg B, Romdhani S and Vetter T. 2009. A 3D face model for pose and illumination invariant face recognition//Proceedings of the 6th IEEE International Conference on Advanced Video and Signal based Surveillance. Genova, Italy: IEEE: 296-301 [DOI: 10.1109/AVSS.2009.58]

Peebles W, Peebles J, Zhu J Y, Efros A and Torralba A. 2020. The hessian penalty: a weak prior for unsupervised disentanglement//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 581-597 [DOI: 10.1007/978-3-030-58539-6_35]

Pei C H, Wu F P, Huang L Q and Zhuang X H. 2021. Disentangle domain features for cross-modality cardiac image segmentation// Medical Image Analysis. 71: #102078 [DOI: 10.1016/j. media. 2021.102078]

Peng X, Yu X, Sohn K, Metaxas D N and Chandraker M. 2017. Reconstruction-based disentanglement for pose-invariant face recognition//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 1632-1641 [DOI: 10.1109/ICCV.2017.180]

Peng X C, Huang Z J, Sun X M and Saenko K. 2019. Domain agnostic learning with disentangled representations//Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA: PMLR: 5102-5112

Pu N, Chen W, Liu Y, Bakker E M and Lew M S. 2020. Dual Gaussian-based variational subspace disentanglement for visible-infrared person re-identification//Proceedings of the 28th ACM International Conference on Multimedia. Seattle, USA: ACM: 2149-2158 [DOI: 10.1145/3394171.3413673]

Reed S, Sohn K, Zhang Y T and Lee H. 2014. Learning to disentangle factors of variation with manifold interaction//Proceedings of the 31st International Conference on Machine Learning. Beijing, China: JMLR.org: 1431-1439

Rezende D J, Mohamed S and Wierstra D. 2014. Stochastic backpropagation and approximate inference in deep generative models//Proceedings of the 31st International Conference on Machine Learning. Beijing, China: JMLR.org: 1278-1286

Rezende D J and Mohamed S. 2015. Variational inference with normalizing flows//Proceedings of the 32nd International Conference on Machine Learning. Lille, France: JMLR.org: 1530-1538

Ridgeway K and Mozer M C. 2018. Learning deep disentangled embeddings with the f-statistic loss//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montreal, Canada: Curran Associates Inc.: 185-194

Roberson P L, McLaughlin P W, Narayana V, Troyer S, Hixson G V and Kessler M L. 2005. Use and uncertainties of mutual information for computed tomography/magnetic resonance (CT/MR) registration post permanent implant of the prostate//Medical physics, 32(2): 473-482

Ruan D L, Yan Y, Chen S, Xue J H and Wang H Z. 2020. Deep disturbance-disentangled learning for facial expression recognition// Proceedings of the 28th ACM International Conference on Multimedia. Seattle, USA: ACM: 2833-2841 [DOI: 10.1145/3394171.3413907]

Sanchez E H, Serrurier M and Ortner M. 2020. Learning disentangled representations via mutual information estimation//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 205-221 [DOI: 10.1007/978-3-030-58542-6_13]

Sangkloy P, Burnell N, Ham C and Hays James. 2016. The sketchy database: learning to retrieve badly drawn bunnies//ACM Transactions on Graphics (TOG). 35(4): 1-12 [DOI: 10.1145/2897824.2925954]

Schuldt C, Laptev I and Caputo B. 2004. Recognizing human actions: a local SVM approach//Proceedings of the 17th International Conference on Pattern Recognition. Cambridge, UK: IEEE: 32-36 [DOI: 10.1109/ICPR.2004.1334462]

Shen Z Q, Huang M Y, Shi J P, Xue X Y and Huang T S. 2019. Towards instance-level image-to-image translation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 3683-3692

Singh K K, Ojha U and Lee Y J. 2019. FineGAN: unsupervised hierarchical disentanglement for fine-grained object generation and discovery//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 6483-6492 [DOI: 10.1109/CVPR.2019.00665]

Sønderby C K, Raiko T, Maaløe L, Sønderby S K and Winther O. 2016. Ladder variational autoencoders//Proceedings of the 30th Interna-

tional Conference on Neural Information Processing Systems. Barcelona, Spain: Curran Associates Inc.: 3745-3753

Soomro K, Zamir A R and Shah M. 2012. UCF101: a dataset of 101 human actions classes from videos in the wild [EB/OL]. [2022-01-21]. https://arxiv.org/pdf/1212.0402.pdf

Sorrenson P, Rother C and Köthe U. 2020. Disentanglement by nonlinear ICA with general incompressible-flow networks (GIN) [EB/OL]. [2022-01-21]. https://arxiv.org/pdf/2001.04872.pdf

Srivastava N, Mansimov E and Salakhutdinov R. 2015. Unsupervised learning of video representations using LSTMs//Proceedings of the 32nd International Conference on Machine Learning. Lille, France: JMLR.org: 843-852

Sun H L, Mehta R, Zhou H, Huang Z C, Johnson S, Prabhakaran V and Singh V. 2019a. DUAL-GLOW: conditional flow-based generative model for modality transfer//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 10610-10619 [DOI: 10.1109/ICCV.2019.01071]

Sun Y, Ye Y, Liu W, Gao W P, Fu Y L and Mei T. 2019b. Human mesh recovery from monocular images via a skeleton-disentangled representation//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 5348-5357 [DOI: 10.1109/ICCV.2019.00545]

Tong B, Wang C, Klinkigt M, Kobayashi Y and Nonaka Y. 2019. Hierarchical disentanglement of discriminative latent features for zero-shot learning//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 11459-11468 [DOI: 10.1109/CVPR.2019.01173]

Tran L, Yin X and Liu X M. 2017. Disentangled representation learning GAN for pose-invariant face recognition//Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA: IEEE: 1283-1292 [DOI: 10.1109/CVPR.2017.141]

Tsai Y H H, Liang P P, Zadeh A, Morency L P and Salakhutdinov R. 2019. Learning factorized multimodal representations [EB/OL]. [2022-01-21]. https://arxiv.org/pdf/1806.06176.pdf

Tulyakov S, Liu M Y, Yang X D and Kautz J. 2018. MoCoGAN: decomposing motion and content for video generation//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 1526-1535 [DOI: 10.1109/CVPR.2018.00165]

Van Steenkiste S, Locatello F, Schmidhuber J and Bachem O. 2019. Are disentangled representations helpful for abstract visual reasoning?//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates, Inc.: 14245-14258

Wah C, Branson S, Welinder P, Perona P and Belongie S. 2011. The Caltech-UCSD Birds-200-2011 Dataset. California Institute of Technology

Wang G Q, Han H, Shan S G and Chen X L. 2020a. Cross-domain face presentation attack detection via multi-domain disentangled representation learning//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 6677-6686 [DOI: 10.1109/CVPR42600.2020.00671]

Wang H, Deng C, Liu T and Tao D. 2021. Transferable coupled network for zero-shot sketch-based image retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence. 44(12): 9181-9194 [DOI: 10.1109/TPAMI.2021.3123315]

Wang W J, Shi Y F, Chen S M, Peng Q M, Zheng F and You X G. 2021. Norm-guided adaptive visual embedding for zero-shot sketch-based image retrieval//Proceedings of the 30th International Joint Conference on Artificial Intelligence. 2021: 1106-1112 [DOI: 10.24963/ijcai.2021/153]

Wang Y H, Bilinski P, Bremond F and Dantcheva A. 2020b. G3AN: disentangling appearance and motion for video generation//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 5263-5272 [DOI: 10.1109/CVPR42600.2020.00531]

Wei L, Zhang S, Gao W and Tian Q. 2018. Person transfer GAN to bridge domain gap for person re-identification//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 79-88

Wu R L and Lu S J. 2020. LEED: label-free expression editing via disentanglement//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 781-798 [DOI: 10.1007/978-3-030-58610-2_46]

Wu S, Deng G C, Li J C, Li R, Yu Z W and Wong H S. 2019. Enhancing TripleGAN for semi-supervised conditional instance synthesis and classification//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 10091-10100

Xiao F Y, Liu H T and Lee Y J. 2019a. Identity from here, pose from there: self-supervised disentanglement and generation of objects using unlabeled videos//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 7012-7021 [DOI: 10.1109/ICCV.2019.00711]

Xiao H, Rasul K and Vollgraf R. 2017. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms [EB/OL]. [2022-01-21]. https://arxiv.org/pdf/1708.07747.pdf

Xiao J, Liao L, Liu Q G and Hu R M. 2019b. CISI-net: explicit latent content inference and imitated style rendering for image inpainting. Proceedings of 2019 AAAI Conference on Artificial Intelligence, 33(1): 354-362 [DOI: 10.1609/aaai.v33i01.3301354]

Xu X X, Yang M L, Yang Y H and Wang H. 2021. Progressive domain-independent feature decomposition network for zero-shot sketch-based image retrieval//Proceedings of the 29th International Joint Conference on Artificial Intelligence. Yokohama, Japan: IJCAI.org: 984-990

Xuan S Y and Zhang S L. 2021. Intra-inter camera similarity for unsupervised person re-identification//Proceedings of 2021 IEEE Confer-

ence on Computer Vision and Pattern Recognition. IEEE: 11926-11935

Yang J L, Dvornek N C, Zhang F, Chapiro J, Lin M D and Duncan J S. 2019. Unsupervised domain adaptation via disentangled representations: application to cross-modality liver segmentation//Proceedings of the 22nd International Conference on Medical Image Computing and Computer Assisted Intervention. Shenzhen, China: Springer: 255-263 [DOI: 10.1007/978-3-030-32245-8_29]

Yang J M, Reed S, Yang M H and Lee H. 2015. Weakly-supervised disentangling with recurrent transformations for 3D view synthesis//Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press: 1099-1107

Yang L L and Yao A. 2019. Disentangling latent hands for image synthesis and pose estimation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 9869-9878 [DOI: 10.1109/CVPR.2019.01011]

Yin G J, Liu B, Sheng L, Yu N H, Wang X G and Shao J. 2019. Semantics disentangling for text-to-image generation//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 2322-2331 [DOI: 10.1109/CVPR.2019.00243]

Yu X M, Chen Y Q, Li T, Liu S and Li G. 2019. Multi-mapping image-to-image translation via learning disentanglement//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates, Inc.: 2994-3004

Yu X M, Ying Z Q, Li T, Liu S and Li G. 2018. Multi-mapping image-to-image translation with central biasing normalization [EB/OL]. [2022-01-21]. https://arxiv.org/pdf/1806.10050.pdf

Zang X H, Li G, Gao W and Shu X J. 2021. Learning to disentangle scenes for person re-identification. Image and Vision Computing. 116: #104330 [DOI: 10.1016/j.imavis.2021.104330]

Zhang J F, Huang Y Y, Li Y Y, Zhao W J and Zhang L Q. 2019a. Multi-attribute transfer via disentangled representation. Proceedings of 2019 AAAI Conference on Artificial Intelligence, 33(1): 9195-9202 [DOI: 10.1609/aaai.v33i01.33019195]

Zhang K Y, Yao T P, Zhang J, Tai Y, Ding S H, Li J L, Huang F Y, Song H C and Ma L Z. 2020. Face anti-spoofing via disentangled representation learning//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 641-657 [DOI: 10.1007/978-3-030-58529-7_38]

Zhang Z Y, Tran L, Yin X, Atoum Y, Liu X M, Wan J and Wang N X. 2019b. Gait recognition via disentangled representation learning//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 4705-4714 [DOI: 10.1109/CVPR.2019.00484]

Zhao J, Cheng Y, Cheng Y, Yang Y, Zhao F, Li J S, Liu H Z, Yan S C and Feng J S. 2019. Look across elapse: disentangled representation learning and photorealistic cross-age face synthesis for age-invariant face recognition. Proceedings of 2019 AAAI Conference on Artificial Intelligence, 33(1): 9251-9258 [DOI: 10.1609/aaai.v33i01.33019251]

Zhao S J, Song J M and Ermon S. 2017. Learning hierarchical features from deep generative models//Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia: JMLR.org: 4091-4099

Zhao Y, Xiong Y J and Lin D H. 2018. Recognize actions by disentangling components of dynamics//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 6566-6575 [DOI: 10.1109/CVPR.2018.00687]

Zheng L, Shen L, Tian L, Wang S, Wang J and Tian Q. 2015. Scalable person re-identification: a benchmark//Proceedings of 2015 IEEE International Conference on Computer Vision (ICCV). Santiago, Chile: IEEE: 1116-1124

Zheng Z, Zheng L and Yang Y. 2017. Unlabeled samples generated by GAN improve the person re-identification baseline in vitro//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 3754-3762

Zheng Z, Yang X, Yu Z, Zheng L, Yang Y and Kautz J. 2019. Joint discriminative and generative learning for person re-identification//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE: 2138-2147

Zhu J Y, Park T, Isola P and Efros A A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 2242-2251 [DOI: 10.1109/ICCV.2017.244]

Zhu J Y, Zhang Z T, Zhang C K, Wu J J, Torralba A, Tenenbaum J B and Freeman W T. 2018. Visual object networks: image generation with disentangled 3D representation//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montreal, Canada: Curran Associates Inc.: 118-129

Zhu X Q, Xu C and Tao D C. 2020a. Learning disentangled representations with latent variation predictability//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 684-700 [DOI: 10.1007/978-3-030-58607-2_40]

Zhu Y Z, Min M R, Kadav A and Graf H P. 2020b. S3VAE: self-supervised sequential VAE for representation disentanglement and data generation//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 6537-6546 [DOI: 10.1109/CVPR42600.2020.00657]

Zhu Z Y, Luo P, Wang X G and Tang X O. 2014. Multi-view perceptron: a deep model for learning face identity and view representations//Proceedings of the 27th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press: 217-225

Zou Y, Yang X D, Yu Z D, Vijaya Kumar B V K and Kautz J. 2020. Joint disentangling and adaptation for cross-domain person re-identification//Proceedings of the 16th European Conference on

Computer Vision. Glasgow, UK: Springer: 87-104 [DOI: 10.1007/978-3-030-58536-5_6]

Zwicker M, Hu Q Y, Szabó A, Portenier T and Favaro P. 2018. Disentangling factors of variation by mixing them//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 3399-3407 [DOI: 10.1109/CVPR.2018.00358]

作者简介

李雅婷，女，硕士研究生，主要研究方向为图像处理和计算机视觉。E-mail: greenallee@whu.edu.cn

肖晶，通信作者，女，副教授，主要研究方向为图像/视频处理和压缩与分析。E-mail: jing@whu.edu.cn

廖良，男，研究员，主要研究方向为图像处理和传输。E-mail: liang@nii.ac.jp

王正，男，教授，博士生导师，主要研究方向为多媒体内容分析。E-mail: wangzwhu@whu.edu.cn

陈文益，男，硕士研究生，主要研究方向为自动驾驶场景的语义分割、计算机视觉和人工智能。E-mail: wenyichen@whu.edu.cn

王密，男，教授，博士生导师，主要研究方向为可测量的无缝立体正射影像数据库、地理信息系统以及全球导航卫星系统、遥感和GIS的集成。E-mail: wangmi@whu.edu.cn