引用格式：毛琪，陈澜．基于解耦内容–风格特征表示的图像转换研究进展[J]．中国传媒大学学报（自然科学版），2023，30(02)：17-30.
文章编号：1673-4793(2023)02-0017-14

基于解耦内容–风格特征表示的图像转换研究进展

毛琪*，陈澜
（中国传媒大学信息与通信工程学院，北京 100024）

摘要：图像到图像转换（Image-to-Image Translation, I2IT）是在保留图像的内容特征条件下，将源域图像转换成目标域图像的过程，旨在学习不同域图像之间的映射。因I2IT应用广泛，如图像风格迁移、图像语义分割、图像修复和图像超分辨率等，图像转换任务一直是计算机视觉领域中研究的热点和重点，并在近年来因深度学习的蓬勃发展取得了显著进展。其中，基于解耦内容–风格特征表示的无监督模型是图像转换的重要方法。本文从内容表示和风格表示两方面梳理了此类模型的发展历程；总结了图像转换任务中常用的数据集和评价指标，同时比较了经典模型在不同数据集上的效果；最后对基于解耦内容–风格特征表示的无监督图像转换的研究进行总结与展望。

关键词：图像到图像转换；隐空间解耦；生成对抗网络
中图分类号：TP391.4　**文献标识码**：A

An overview of disentangled content-style representation based image-to-image translation

MAO Qi *, CHEN Lan
(Communication University of China, Beijing100024, China)

Abstract: Image-to-Image Translation (I2IT) aims to learn the mapping between images of different domains. Image translation is a hot topic in computer vision since I2IT models can be applied in various fields, including image style transfer, semantic segmentation, and image super-resolution. With the booming development of deep learning, numerous effective I2IT models have emerged in recent years. Among them, unsupervised models based on disentangled content-style representation are essential methods. In this paper, we first review the development of such models in terms of content and style representation. Then, we summarize the common datasets and metrics used in I2IT tasks and compare the results of various models. Finally, we assess and forecast future trends in I2IT development.

Keywords:GAN; image-to-image translation; disentangled latent space

1　引言

唤醒黑白相片，还原历史色彩；重绘自然风景，打造艺术世界；勾勒动漫形象，打破“次元”壁垒……得益于计算机视觉的发展，各种智能图像处理应用丰富和便利了人们的日常生活。早期，不同的图像处理任务分别由特定的模型单独完成，Isola等人[1]在2016年首次提出图像到图像转换的概念（Image-to-Image translation, I2I），其目标是将图像从一个图像域转换到另一个图像域，其中图像域定义为共享视觉特征的

基金项目：中国传媒大学国家重点实验室专项项目（CUC22GZ035）；国家自然科学青年基金项目（62201522）
作者简介（*为通讯作者）：毛琪（1995-），女，讲师，主要从事图像、视频生成研究。Email：qimao@cuc.edu.cn；陈澜（2001-），女，本科生，主要从事图像、视频生成研究。Email：eva_cl@cuc.edu.cn

一类图像。转换过程期望图像的风格特征满足目标图像域的分布,而图像本身的内容特征保持不变。如图1所示,统一的图像转换模型可以解决包括语义图像合成[2,3]、图像增强[4]和风格迁移[5,6]等任务,是当前图像合成领域研究的热点之一。

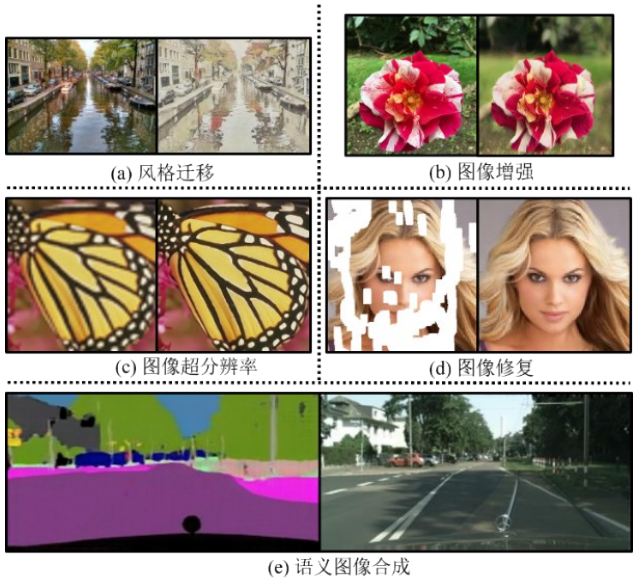

图1 图像转换应用举例

与图像转换任务最具有相关性的任务是风格迁移[7],如图2(a)所示,风格迁移后的图像期望能够保持源图像的内容,迁移参考图像的风格。其中内容通常指输入图像的结构信息,风格通常指参考图像的纹理与颜色信息;而图像转换中内容和风格的含义不固定,与数据集本身的特性有关。如图2(b)所示,当源图像域是风景照片,目标图像域为艺术风格画时,内容特征和风格特征与风格迁移任务相同;当源图像域是男性,目标图像域是女性时,内容特征是姿态、五官和脸型,而风格特征是头发、妆容等;当源图像域是猫,目标图像域是狗时,内容特征是姿势、朝向和表情,风格特征是颜色、形状、毛发等。由此可见,图像转换任务的定义更加宽泛,其风格和内容的具体含义是通过数据集的学习得到的。一个更准确的定义是,内容特征指图像转换中的域不变(Domain-Invariant)特征,风格特征指图像转换中的域特定(Domain-Specific)特征。

早期的图像转换模型[1,4]直接通过生成模型,隐式建模内容和风格特征的变换,转换过程中保持输入图像内容特征不变,改变风格特征,给定输入图像只能得到唯一的输出结果,极大地限制了图像转换过程中的可控性和多样性。为了解决这个问题,Huang等

图2 神经风格迁移(左)与图像转换(右)对比示例

人[8]和Lee等人[9]首次提出了基于解耦内容-风格特征表示的图像转换模型,对不同图像域的风格和内容特征显示建模,从而能够实现基于样例图像引导和基于随机风格向量引导的多样且可控的图像转换,后续研究者们也沿着这个思想进行更深入地探索。

尽管目前的研究已经取得了很大的成功,但图像转换仍然存在很多未被清晰定义的问题,例如,如何构建一个更好的风格和内容特征的表示,如何更准确地评价图像转换的结果等。为了进一步挖掘模型的潜力,探究未来的改进方向,本文对基于解耦内容-风格特征表示的图像转换模型的研究现状和进展进行综述。本文的第2节首先对图像转换目前的主要研究问题和基于解耦内容-风格特征表示模型的基本框架进行简要介绍;第3节对基于解耦内容-风格特征表示模型的发展脉络以及研究现状进行梳理;第4节对图像转换任务中常见数据集和评价指标进行整理与归类,并对经典模型进行定量与定性的对比和评价;第5节总结了此类模型的发展历程,并对未来的发展方向进行思考和展望。

2 图像转换概述

图像转换的核心在于学习不同图像域之间的映射,这与生成模型有很高的相关性。生成模型使用特定网络结构来建模一类图像的分布,从而可以采样生成类似于样本数据、服从同一分布的图像。同样地,在图像转换中,生成模型以损失函数为约束条件进行映射学习,使输出符合目标域图像的分布。在图像转换领域,变分自编码器[25](Variational Auto Encoder,VAE)和生成对抗网络[26](Generative Adversarial Network,GAN)是最常用且最有效的生成模型。VAE通过最大化对数似然下限来模拟数据分布,GAN则试图寻找生成器和鉴别器之间的纳什平衡。

经过近些年的发展,图像转换模型从有监督到无监督(图3a)、从一对一映射到一对多映射(图

3c）、从双域到多域（图3b），逐渐走向成熟和完善。表1对部分经典的图像转换模型进行了梳理和归纳。最初，Isola等人[1]提出的Pix2Pix模型使用配对数据集建模映射函数，BicycleGAN[10]在此基础上进行改进，使模型具有一对多映射的能力。由于构建不同域的配对图像数据集难度大、代价高，CycleGAN[4]、DiscoGAN[27]等模型提出循环一致性约束，在非配对数据集来建立域间的双向关系。为了进一步使模型同时具有无监督训练和多输出的能力，Huang等人[8]和Lee等人[9]提出了基于解耦内容-风格特征表示的图像转换模型，开拓了图像转换模型的一个全新结构分支。

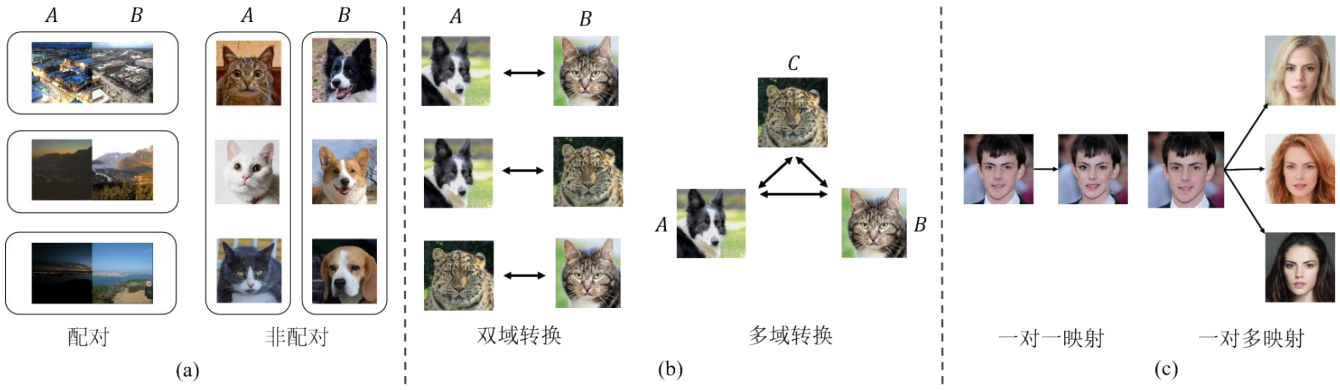

图3 图像转换模型分类

表1 图像到图像转换经典模型概览

分类	模型	多样性	多域	内容-风格解耦	年份
有监督	Pix2Pix[1]				2017
	BicycleGAN[10]	√			2017
	Pix2PixHD[2]				2018
	SPADE[3]				2019
无监督	UNIT[11]				2017
	CycleGAN[4]				2017
	MUNIT[8]	√		√	2018
	DRIT[9]	√		√	2018
	U-GAT-IT[12]				2019
	FUNIT[13]		√	√	2019
	DMIT[14]	√	√	√	2019
	TransGaGa[15]			√	2019
	DRIT++[5]	√	√	√	2020
	DSMAP[16]	√		√	2020
	StarGAN-v2[17]	√	√	√	2020
	Swap AE[18]			√	2020
	TSIT[19]			√	2020
	TUNIT[20]		√	√	2021
	HCSI2I[21]	√	√	√	2021
	SAVI2I[22]	√	√	√	2022
	SA-Dis[23]	√	√	√	2022
	GP-UNIT[24]			√	2022

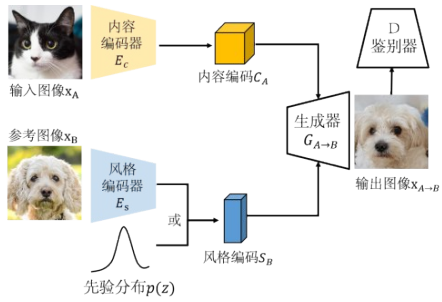

图4 基于解耦内容-风格特征表示的图像转换模型

当前图像转换的关键问题在于无监督的多域转换和一对多映射。从表1可以看出,基于解耦内容-风格特征表示模型是有效且高效的解决方式。如图4所示,在此模型中,图像被嵌入到两个隐空间:域间共享的内容空间和域内特定的风格(样式、属性)空间,分别用内容编码和风格编码表示。

具体而言,此类模型包含风格编码器、内容编码器、生成器和鉴别器。如图4所示,以图像域A到图像域B的转换为例,生成器利用内容编码c_A和风格编码s_B,生成属于B域的图像$x_{A\to B}$。其中,c_A来自内容编码器E_c,s_B根据模型的不同,来自风格编码器E_s或先验分布$p(z)$。表达式如公式(1)所示:

$$\begin{aligned} x_{A\to B} &= G(c_A, s_B) \\ c_A &= E_c(x_A), s_B = E_s(x_B) \; or \; p(z). \end{aligned} \tag{1}$$

3 基于解耦内容-风格特征表示的图像转换算法研究现状

从解耦的角度出发,内容特征和风格特征的提取与空间的构建是图像转换的关键,也是本文研究的重点。本文3.1节和3.2节分别梳理了此类模型在风格特征建模和内容特征建模上的改进与发展。

3.1 风格特征建模

直观上,域特定空间中的风格编码应具有多样性、灵活性与可控性。以此为目的,风格空间的发展(图5)主要从空间构建的角度出发,从双域到多域、从分离到统一,取得了极大的进展。

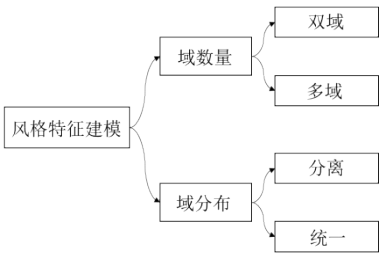

图5 风格特征的发展脉络

MUNIT[8]和DRIT[9]是解耦内容-风格I2IT模型的开篇之作。其核心有二:一是引入随机风格向量,使模型可以进行一对多的转换。DRIT[9]使用KL损失(公式2)显式地将风格编码嵌入正态分布;MUNIT[8]则将从正态分布中采样的向量作为输入生成器的风格编码,通过重建损失(公式3)使风格编码器的输出和采样得到的向量达到一致。二是引入循环一致性损失,使模型可以利用非配对的图像数据集。

$$\begin{aligned} \mathscr{L}_{KL} &= \mathrm{E}\left[D_{KL}(s \| N(0,1)) \right] \\ D_{KL}(p\|q) &= -\sum p(z) \log \frac{p(z)}{q(z)} dz \end{aligned} \tag{2}$$

$$\mathscr{L}_{sty} = \mathrm{E}_x\left[\| s - E_s(G(E_c(x), s)) \|_1 \right], \mathrm{s}\sim\mathrm{N}(0,1) \tag{3}$$

然而,与图像域一一对应的风格编码器具有局限性,模型无法扩展应用到多域图像转换。为了解决此问题,Yu等人[14]提出DMIT模型,所有图像域共用一个风格编码器。如图6(a)所示,域标签作为风格编码的域标识,与风格编码在通道维度上拼接后输入生成器。但是这种完全忽略域信息的风格编码器在图像域间差异较大时效果不佳。Lee等人[5]在DRIT[9]的基础之上,提出了使用统一风格编码器的DRIT++。如图6(b)所示,通过将域标签和图像共同作为输入,风格编码器可根据域标签提取各域特有的风格特征。

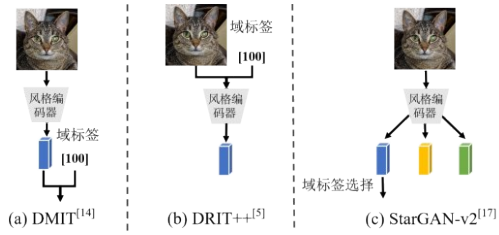

图6 多域转换的风格编码器

和DRIT++[5]不同,StarGAN-v2[17]在风格编码器的输出层采用多分支结构。如图6(c)所示,域标签用于选择对应的输出分支作为图像的风格编码。统一的风格编码器不仅简化了模型结构,还使其在多图像域训练中获益,获得更强的泛化能力。此外,StarGAN-v2[17]增加了同样具有多分支输出结构的映射网络模块。从先验分布采样的随机向量不直接作为风格编码,而是经由此网络被映射到各域的风格空间后,由域标签选择对应维度的输出。

映射网络和风格编码器输出层分支结构的设计显式地分离了不同域的风格编码,使其更准确地捕捉到了域特定的风格特征,产生更多样化的图像。然而,分离的分布使模型不具备域间连续转换的能力。

为了解决此问题，Liu等人[28]在StarGAN-v2[17]的基础上加入了两个和风格特征相关的损失项：公式(4)为三元组损失，其中，α为边距常量，保证各域的风格向量相互分离的同时，控制域间的紧凑程度；s_a、s_p和s_n为风格编码，s_n的所属域不同于s_a和s_p；公式(5)为风格正则化，通过惩罚较大的风格编码l_2范式，风格空间以原点为中心收缩。其中，s表示风格编码。

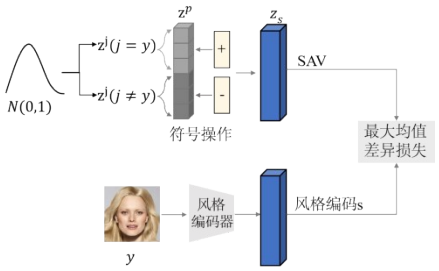

图7 SAVI2I示意图

$$\mathscr{L}_{tri} = \mathrm{E}_{s_a s_p s_n}[\max(\|s_a - s_p\| - \|s_a - s_n\| + \alpha, 0)] \quad (4)$$
$$s_a, s_p \in S_i, s_n \in S_j, j \neq i$$

$$\mathscr{L}_{SR} = \mathrm{E}_s[\|s\|_2^2] \quad (5)$$

Mao等人[22]则从编码的角度出发，提出域共享的统一风格空间，利用符号操作对图像域信息进行编码，使得域间插值的风格编码能够产生合理的结果。如图7所示，首先从高斯分布中采样d×N维向量z^p，其中N表示域的数量，d为每个域风格属性向量的维度。其次根据域标签构造有符号向量(Signed Attribute Vector，SAV)z_s。然后使用最大均值差异(公式8)统一z_s和风格编码器的输出z。z_s和z^p的计算公式如公式(6)、公式(7)所示：

$$z_s = O_y(z^p) \quad z^p \in \mathbb{R}^{d*N}, z^p \sim N(0, I), y \in \{1...N\} \quad (6)$$

$$O_y(z^p) = [-|z_1^1|, -|z_2^1|, ..., -|z_d^1|, ..., +|z_1^y|, +|z_2^y|, ..., +|z_d^y|, ..., -|z_1^N|, -|z_2^N|, ..., -|z_d^N|] \quad (7)$$

$$\mathscr{L}_{MMD} = \mathrm{E}_{p(z^s)p(z)}[k(z^s, z)] + \mathrm{E}_{q(z^s)q(z)}[k(z^s, z)] - 2\mathrm{E}_{p(z^s)q(z)}[k(z^s, z)], k(z^s, z) = e^{-\frac{\|z^s - z\|}{2\sigma^2}} \quad (8)$$

上述使用域标签的无监督I2IT模型已极大地降低了数据集的收集难度。但是，当数据集非常庞大时(如FFHQ[29])，为每一张图片标记域信息同样成本高昂，且对于有些数据集来说，域的划分是多样且模糊的。针对此问题，Baek等人[20]提出了TUNIT模型，创造性地引入自监督训练的思想。过MI[30]和InfoNCE[31]等约束，风格编码器能够充分利用域特定的风格信息，自动判断输入图像的所属域，输出伪标签。

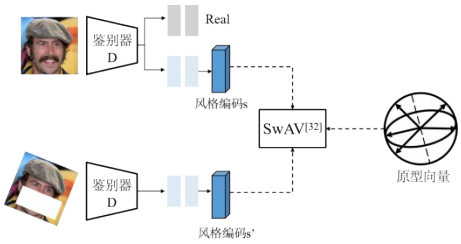

图8 原型向量示意图

然而，TUNIT[20]无法实现多样式输出，并且存在错误分类的问题。Kim等人[23]指出，这种错误产生的原因在于单一的域标签区分方式并未考虑到域间的语义距离。同时，域标签将I2IT限制在预先定义的图像域中，无法控制训练所用标签之外的域。为了解决此问题，Kim等人[23]使用一组标准化的原型向量来统一风格空间，每个原型向量可以被简单地理解为各域风格编码的聚类中心。此外，风格编码器被集成到鉴别器中，共享骨干网络。如图8所示，通过SwAV[32]聚类方法，风格编码被嵌入原型向量空间，在摆脱域标签约束的同时，模型可以采样原型向量，产生多样输出。

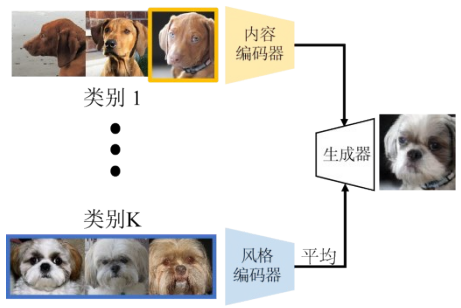

图9 FUNIT[13]训练示意图

虽然上述模型取得了良好的转换效果，但其训练阶段需依赖大量的图像数据，且无法基于先验知识从少量样本中获得泛化能力，应用于不属于训练数据的图像域。针对此问题，Liu等人[13]提出了少样本学习模型FUNIT。如图9所示，训练时使用包含K个类别的图像数据集S，对其中某一类的一张图像进行内容编码，对另一类的一组图像分别进行风格编码后求算术平均。测试时通过少量不属于S的新目标域图像作为风格指导，模型就能将S中的任意一类图像转换到新目标域。

3.2 内容特征建模

如图10所示，内容编码器的发展从信息获取的角度出发，通过增加内容映射模块，内容特征的空间分

布从域共享到域特定;通过引入生成先验等来增强不同语义级别的特征提取能力和表现能力,不同域的内容空间关系从语义对应扩展到语义非对应。

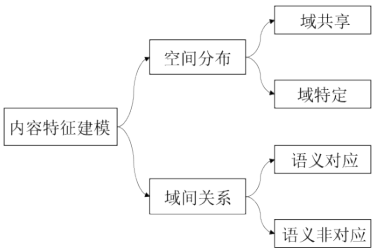

图10 内容特征的发展脉络

MUNIT[8]和DRIT[9]中包含两个和图像域一一对应的内容编码器。为了统一各域的内容空间,保证内容分布一致,DRIT[9]共享两个内容编码器的最后一层和两个生成器的第一层权重,并提出了内容对抗损失(公式9);MUNIT[8]则证明了通过图像转换映射的学习,两个域的内容编码分布在隐式空间达到一致。但是,域间完全共享内容空间的假设损失了部分图像特有的内容信息,降低了内容编码的表达能力。因此,Chang等人[16]提出DSMAP,将共享内容空间的特征二次映射到域特定的空间中,使图像的内容信息得到更充分的表达。如图11所示,内容编码器输出共享空间的内容特征后,由映射函数φ_B将其二次投影至目标域B的内容空间,得到内容编码$C_{A\to B}$。映射函数φ_A则将内容特征重映射至图像原属域A,得到$C_{A\to A}$。φ_A映射通过内容编码器中间层特征和$C_{A\to A}$组成的域特定内容损失学习。

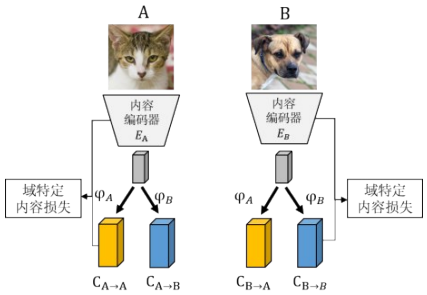

图11 DSMAP域特定内容空间示意图

上述模型在语义对应且几何形状差距不大的图像域上(如猫和狗、男性和女性等)取得了良好的转换效果,但无法应用于语义相似但空间分布差异较大的情况(图13)。为了解决此问题,Wu等人[15]提出TransGaGa模型,如图12所示,通过几何估计器和几何转换模块提炼、映射图像的几何结构信息,内容编码能够学习图像的高级语义表示。

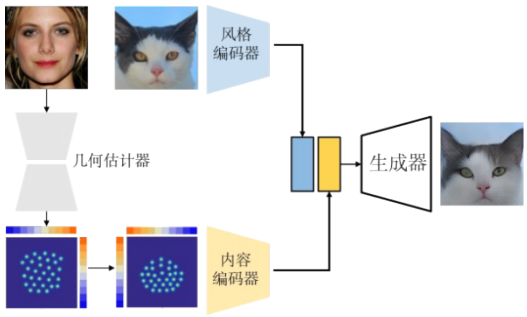

图12 TransGaGa[15]部分模型结构图

为了进一步将图像转换扩展到无语义对应的数据集(如轿车和鸟类),Yang等人[24]提出GP-UNIT模型,通过引入BigGAN[34]生成先验,内容编码器能够学习更为抽象的内容信息,如方位、布局等,并建立域间对应关系。图14展示了训练的第一阶段,首先采样噪声,利用BigGAN[34]生成两张具有相似方位和布局的不同类图像。其次,内容编码器将两张图像分别编码为单通道灰度图以消除域信息。最后,将单通道灰度图输入解码器F。F的第I部分预测输入图像的形状,第II部分重建输入图像。如图13所示,得益于强大的内容编码器,GP-UNIT[24]成功实现了异质、不对称的图像转换。

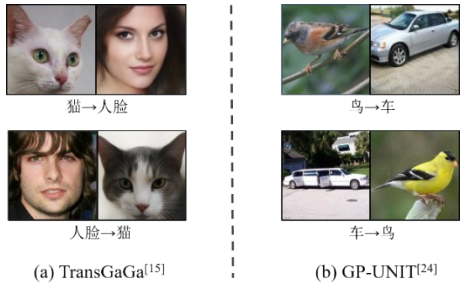

图13 TransGaGa[15](左)与GP-UNIT[24](右)的转换效果

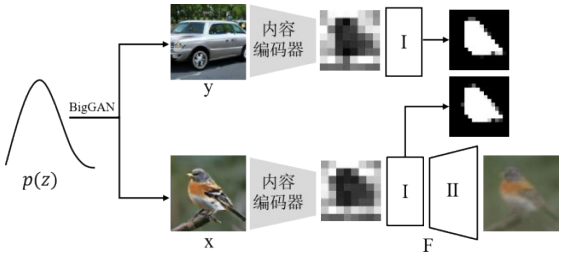

图14 GP-UNIT[24]第一阶段训练示意图

3.3 内容-风格的解耦与融合

以内容-风格解耦为前提,上述模型从风格和内容两方面进行改进和完善。而如何强制风格特征和

内容特征被分别提取和利用以满足前提，是模型运转的关键。

损失函数的约束是解耦的基础。根据域共享内容空间和内容信息在转换中保持不变的假设，Lee等人[9]提出了内容对抗损失（公式9）和跨域循环一致损失（公式15）、Liu等人[28]提出内容保留损失（公式10）。从风格表示多样的特点出发，为了缓解模式崩溃问题，Mao等人[35]提出模式查找正则化（公式11）、Baek等人[20]引入风格对比损失（公式12）。而Choi等人[36]认为，模式崩溃的一个重要原因是单一的决策边界，而这两种损失函数并不能从根本上解决此问题。因此，Choi等人[36]从GAN的基本原理出发，引入灵活判决边界机制，提出样式引导鉴别器损失（公式14）。此外，规范化点互信息（公式13）用于消除潜在空间中编码风格表征的纠缠，从而进一步缓解模式崩溃。

风格特征和内容特征在生成器中通常使用AdaIN[37]进行融合，也有一些模型[23,38]使用目前最先进的生成模型StyleGAN[29]。生成对抗损失（公式16）保证生成器将风格编码和内容编码相结合，输出真实且属于目标域的图像。为了进一步让鉴别器接近纳什平衡点，生成更逼真的图像，Choi等人[36]使用重要性采样，如公式（20）所示，根据鉴别器的输出，将特定权重分配给生成器。此外，风格编码重建损失（公式17）和内容编码重建损失（公式18、19）用于促进图像和隐空间的逆映射，强制生成器在生成图像时利用风格编码和内容编码。

3.4 存在的问题

目前，基于解耦内容-风格特征表示的图像转换模型侧重于探索内容和风格的空间分布，缺少对风格和内容的具体内涵的研究。图像表达的特征包罗万象，颜色、纹理、形状、结构和语义等方面均蕴含大量可变性。对于图像转换模型来说，最关键的核心就是捕捉并且改变某些特征。解耦内容-风格特征表示模型的风格编码与内容编码控制的是哪些特征、控制的程度能达到多少等问题值得深入探讨。

风格和内容在形式上是解耦的，但在含义上是关联互补的，模型需要根据图像数据的不同权衡风格和内容特征的分配，并调整其变化程度。因此，为了更好地研究此类图像转换模型在不同数据集上对内容和风格特征的学习能力和适应程度，本文第4节根据域间差异性对图像转换常用数据集进行进归类和对比。

表2 基于解耦内容–风格图像转换模型的常用损失函数

分类	名称	公式	
解耦	内容对抗损失	$\mathcal{L}_{adv}^{content}(E,D)=\mathrm{E}[\frac{1}{2}\log D^{c}(E(x_i))+\frac{1}{2}\log(1-D^{c}(E(x_i)))]$ $+\mathrm{E}[\frac{1}{2}\log D^{c}(E(x_j))+\frac{1}{2}\log(1-D^{c}(E(x_j)))], i,j\in A,B,C...,i\neq j$	(9)
	内容保留损失	$\mathcal{L}_{cont}=\mathrm{E}_{x,s}[\phi(x,G(x,s))], \phi:LPIPS$[39]	(10)
	模式查找正则化	$\mathcal{L}_{ms}=\mathrm{E}_{x,s_1,s_2}(\frac{\|G(E_c(x),s_1)-G(E_c(x),s_2)\|_1}{\|s_1-s_2\|_1})$	(11)
	风格对比损失	$\mathcal{L}_{sty1}=\mathrm{E}_{x,\tilde{x}}[-\log\frac{\exp(s\cdot s^{+}/\tau)}{\sum_{i=0}^{N}\exp(s\cdot s_i^{-}/\tau)}], s=E_s(x)$	(12)
	规范化点互信息损失	$\mathcal{L}_{NPMI}=\mathrm{E}_{(x,s)\sim(\mathbb{P},\mathbb{R})}[NPMI(s_1,s_2)]$	(13)
	样式引导鉴别器损失	$\mathcal{L}_{STGD}=\mathrm{E}_{(x_r,z)\sim(\mathbb{P},\mathbb{R})}[\log(f_{sig}(C(x_r)-C(x_g)))]$ $+\mathrm{E}_{(x_r,z)\sim(\mathbb{P},\mathbb{R})}[\log(1-f_{sig}(C(x_g)-C(x_r)))]$	(14)
	跨域循环一致损失	$\mathcal{L}_{cyc}=\mathrm{E}_{x,y}[\|x-G(G(E_c(x),s_y),E_s(x))\|_1]$	(15)
融合	生成对抗损失	$\mathcal{L}_{adv}=\mathrm{E}_{x,y}[\log D(x)]+E_{x,y}[\log(1-D(G(x,s_y)))]$	(16)
	风格编码重建损失	$\mathcal{L}_{sty}=\mathrm{E}_{x}[\|s-E_s(G(E_c(x),E_s(x)))\|_1]$	(17)
	内容编码重建损失	$\mathcal{L}_{cont1}=\mathrm{E}_{x}[\|c-E_c(G(E_c(x),E_s(x)))\|_1]$	(18)
		$\mathcal{L}_{cont2}=\mathrm{E}_{x,c}[\frac{1}{WH}\sum_{i,j}^{W,H}\|c-E_c(G(E_c(x),E_s(x))\|_2^2]$	(19)
	重要性加权损失	$\mathcal{L}_{IG}=\mathrm{E}_{(x_g,z)\sim(\mathbb{Q},\mathbb{R})}[e^{C(x_g)+\varepsilon}\log(1-D(x_g))]$	(20)

4 数据集、评价指标及模型比较

数据集的选择是评价模型性能的基础,如表3所示,根据图像域之间的差异程度,数据集可划分为以下三类:

(1)场景类 此类数据集的域间差异为颜色和纹理,形状和语义特征保持不变。比较有代表性的场景数据集如图15(a)所示,Architectural labels2photo[1]数据集,包含配对的建筑物正面图象和其结构标签图;Summer2winter[4]数据集,由非配对的一组夏季风景图和一组冬季风风景图组成。。

(2)真实对象类 此类数据集图像大多为真实事物,如图15(b)所示,包括不同性别的人脸、不同物种的动物等。域间差异程度较大,有颜色、纹理、形状等,转换难度比场景类高。

(3)艺术风格对象类 此类数据集图像包含一组夸张艺术风格图像,如图15(c)所示,真实人脸和动漫人脸在颜色、纹理、形状和语义特征上差异极大,转换难度最高。

表3 数据集总结

类型	数据集名称	年份	出处	数量	分辨率
场景类	Oxford-102[40]	2008	ICVGIP	8189	(1168~556)×(556~500)
	Standford Cars[41]	2012	CVPR	16185	(7800~101)×(5400~41)
	NYU Depth v2[42]	2012	ECCV	1449	640×480
	Architectural labels2photo[1]	2013	CVPR	506	256×256
	LSUN(church)[43]	2015	arXiv	939835	(1462~341)×(256~256)
	LSUN(bedroom)[43]	2015	arXiv	126527	(7800~101)×(5400~41)
	Cityscape labels2photo[44]	2016	CVPR	15000	2048 ×1024
	Edges2shoes/handbags[1]	2016	CVPR	138767/50025	256×128
	Day2night[1]	2016	CVPR	22397	256×128
	Gta2City[45]	2016	ECCV	24966	(911~256)×(609~256)
	Weather[46]	2017	JVCIR	183798	500×(375~35)
	Artist2photo[4]	2017	ICCV	3401	256×256
	Summer2winter[4]	2017	ICCV	2740	256×256
	Iphone2DSLR_flowers[4]	2017	ICCV	6186	(942~360)×(360~78)
	INIT dataset[47]	2019	CVPR	155529	1208× 1920,3000× 4000
	Day2Timelapse[48]	2020	CVPR	63119	480×320
	BDD100K[49]	2020	CVPR	45338	512×256
	Map2aerial photo[1]	2017	CVPR	3292	600×600
	ADK[50]	2019	CVPR	25562	(2048~130)×(2048,96)
真实对象类	CUHK Face Sketch[51]	2009	PAMI	188	200×250
	Apple2Orange[4]	2017	CVPR	2528	256×256
	Horse2zebra[4]	2017	CVPR	2661	256×256
	Grumpifycat[4]	2017	CVPR	302	256×256
	FFHQ[29]	2018	CVPR	70000	1024×1024
	CelebA-HQ(male2female)[17]	2018	ICLR	30000	1024×1024
	MaskCelebA[52]	2019	CVPR	30000	1024×1024
	AFHQ[17]	2020	CVPR	15000	512 × 512
	MetFace[53]	2020	NIPS	1336	1024×1024
艺术风格对象类	Selfie2anime[12]	2020	ICLR	7000	256 × 256
	Face2anime[54]	2021	arXiv	17796	128 × 128

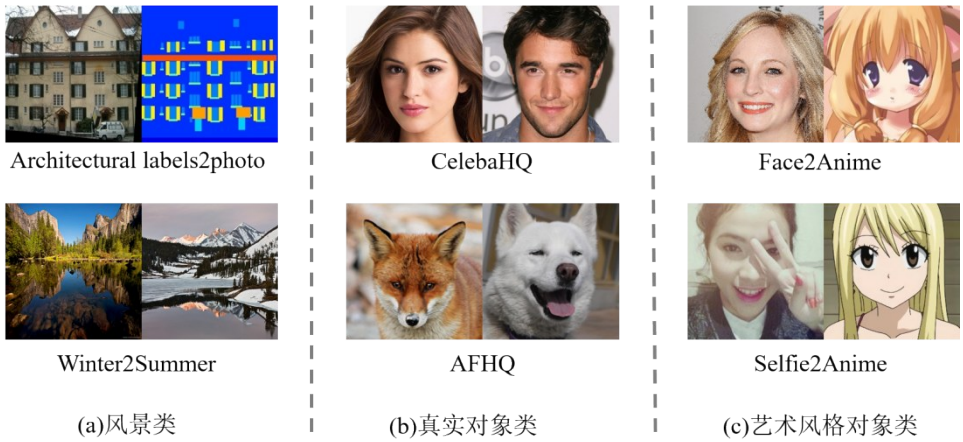

图15 数据集示例

4.2 评价指标

图像转换的评价指标从图像质量、图像多样性、图像在内容上的保持程度以及图像与参考图像在风格上的相似程度这四个方面来衡量模型的性能。以下介绍4种常用指标。

（1）FID(Fréchet Inception Distance)

FID[55]以在ImageNet[56]数据集上训练的Inception-V3[57]模型作为特征提取器，计算真实图片和生成图片的特征向量的距离。计算公式如公式(21)所示。

$$FID(g,r) = \|\mu_g - \mu_r\|_2^2 + Tr(\sum_g + \sum_r - 2(\sum_g \sum_r)^{\frac{1}{2}}) \quad (21)$$

其中，g表示生成图像，r表示真实图像，μ和$\sum$分别表示均值和协方差。当生成图像和真实图像特征的均值和协方差相近时，生成图像的分布接近真实图像的分布，即FID越小，生成的图像质量越好。

（2）LPIPS(Learned Perceptual Image Patch Similarity)

LPIPS[39]度量两张图像的感知距离，其特征提取网络的训练使用真实图像和失真图像，因此LPIPS对真实程度不同的生成图像评估更加鲁棒。计算公式如公式(22)所示：

$$LPIPS(g,r) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \left\| w_l \odot (g_{hw}^l - r_{hw}^l) \right\|_2^2 \quad (22)$$

其中，g和r分别表示生成图像和真实图像，l表示特征提取网络的层数，g_{hw}^l和r_{hw}^l分别表示生成图像和真实图像在第l层输出的特征，H_l和W_l分别表表示第层特征的高度和宽度，w_l表示第一层和第l层特征的余弦距离。LPIPS越小表示两张图像越相似；用于多样性评价时，值越高表示生成的图像越多样。

（3）DIPD(Domain-Invariant Perceptual Distance)

DIPD[13]计算源域图像和转换后图像在VGG[58]网络中Conv5输出特征的距离，衡量转换后图像的内容保持程度。

（4）SIFID(Single Image Fréchet Inception Distance)

SIFID[59]通过计算两幅图像特征之间的FID[55]衡量生成图像和参考图像内部分布的差异。SFID得分越低，表示两张图像风格越相似。

4.3 模型比较

在比较基于解耦内容-风格特征图像转换模型时，通常会分别比较隐向量引导方法和参考图像引导方法。前者指的是通过采样随机向量来生成风格编码，而后者则是利用参考图像来生成风格编码。为了更好地说明风格的学习和内容的保持，本文只比较由参考图像引导的图像转换效果。本小节在Summer2winte[4]、CelebA-HQ[17]、AFHQ[17]和Face2Anime[54]数据集上对部分模型[5,8,16,17,22,23]进行了定性和定量的比较。

4.3.1 定性比较

图16定性比较了不同模型在Summer2winter[4]数据集上的图像转换效果，DRIT++[5]、StarGAN-v2[17]和SAVI2I[22]可以在保持结构特征的同时表现参考图像的风格，而SA-Dis[23]对域不变的内容特征把握不当，过多地改变了输入的结构。

图17定性比较了CelebA-HQ[17]数据集上的图像转换结果，MUNIT[8]和DRIT++[5]仅改变了输入图像的妆容特征，无法改变变化较大的胡须和头发样式；StarGAN-v2[17]和SAVI2I[22]对头发的转换效果较好，较

为准确地捕捉并还原了参考图像中的头发样式;StarGAN-v2[17]对人脸的身份特征保持得最好,但对人脸之外的部分(背景、饰品等)转换效果较差。

图18定性比较了AFHQ[17]数据集的转换效果,MUNIT[8]模型无法在域间差异较大的猫和狗图像域间成功转换;SA-Dis[23]、StarGAN-v2[17]和SAVI2I[22]相比,内容特征(嘴的开合、背景)保持效果以及风格特征(耳朵、鼻子的形状)的改变效果更好。值得注意的是,第四行输入图像的耳朵较大,垂于面部两侧,此特征仅存在于狗的图像中且数量极少,转换难度较高。DSMAP[16]将此特征保留;StarGAN- v2[17]和SAVI2I[22]在此部分产生模糊的结果;DRIT++[5]和SA-Dis[23]较为合理地将输入中的耳朵转换为猫面部的一部分。

图19定性比较了不同模型在Face2Anime[54]数据集上的图像转换效果。DSMAP[16]、StarGAN-v2[17]、SAVI2I[22]和SA-Dis[23]都能够学习参考图像的发色特征,但眼睛的颜色、面部妆容等学习程度不够。对于内容特征,仅有DSMAP[16]和SA-Dis[23]保持了输入图像的方位,而嘴巴的形状、眼睛的张开程度以及视线方向等在所有模型的转换结果中均无法体现。

从模型的角度来看,StarGAN-v2[17]和SAVI2I[22]在四类数据集上的总体转换效果最好,除Face2Anime[54]数据集外,都能较好地保持内容特征、表现风格特征,从具体数据集中学习解耦域不变和域特定的图像特征。

从数据集的角度来看,模型在风景类的Summer2Winter[4]数据集上的整体转换效果最好,在Face2Anime[54]数据集的转换效果最差,艺术对象类数据集对I2I模型仍是一个挑战。

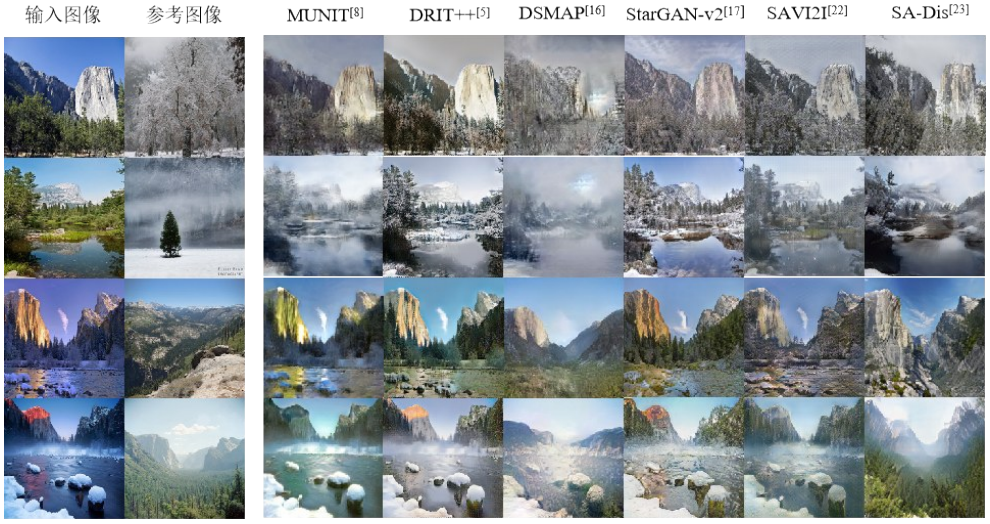

图16 Summer2winter[4]数据集上的定性比较

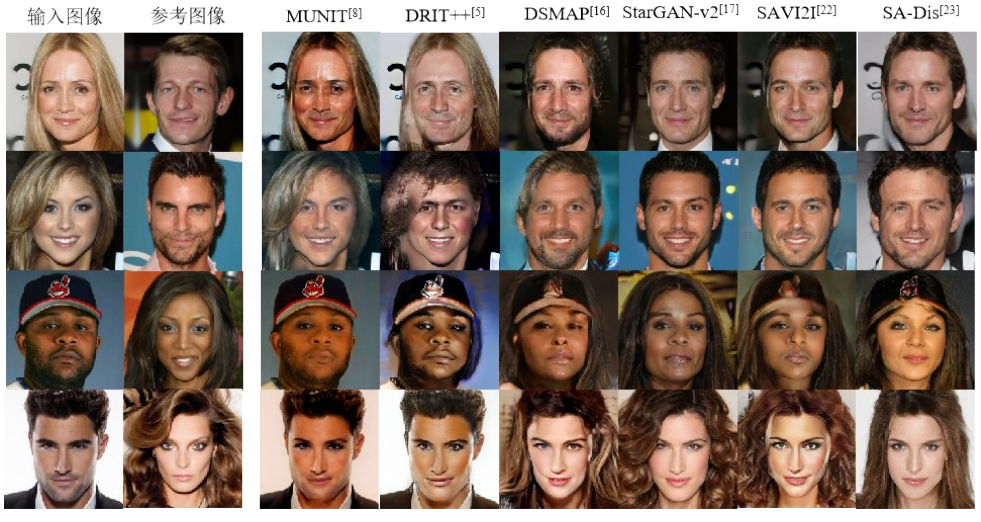

图17 CelebaHQ[17]数据集上的定性比较

图18 AFHQ[17]数据集上的定性比较

图19 Face2anime[54]数据集上的定性比较

表4 模型在不同数据集上的定量比较评价结果

数据集 模型	Summer↔Winter[4]				Male↔Female[17]			
	FID ↓	LPIPS ↑	DIPD ↓	SIFID ↓	FID ↓	LPIPS ↑	DIPD ↓	SIFID ↓
MUNIT[8]	100.9060	0.3325	2.8250	1.2204e−04	64.3030	0.0705	1.0373	4.2361e−05
DRIT++[5]	77.6379	0.1347	2.3593	1.5120e−04	75.8301	0.0998	1.9483	5.7625e−05
DSMAP[16]	97.2408	0.5544	3.1778	1.0235	88.5979	0.2692	2.5232	3.0932e−05
StarGAN−v2[17]	74.4934	0.3152	3.2235	1.3389e−04	55.4247	0.3876	2.9416	1.8231e−05
SAVI2I[22]	79.9192	0.2966	2.9004	4.1119e−05	49.9757	0.3200	2.7313	1.3069e−05
SA−Dis[23]	80.9608	0.4845	3.6538	5.9112e−05	53.3387	0.2722	2.7505	3.8375e−05
数据集 模型	Cat↔Dog[17]				Face↔Anime[54]			
	FID ↓	LPIPS ↑	DIPD ↓	SIFID ↓	FID ↓	LPIPS ↑	DIPD ↓	SIFID ↓
MUNIT[8]	229.5702	0.1881	2.4739	4.0338e−05	160.0509	0.2505	4.1831	5.3623e−05
DRIT++[5]	59.9451	0.6040	3.2690	3.9547e−05	56.5549	0.3734	4.3078	9.5084e−05
DSMAP[16]	68.6781	0.3269	3.3852	1.9594e−05	67.0111	0.3280	3.4690	7.3298e−01
StarGAN−v2[17]	40.7741	0.4278	2.9475	2.0181e−05	55.4247	0.4704	4.2992	5.2856e−05
SAVI2I[22]	37.8581	0.4563	3.1791	1.1716e−05	36.0809	0.4531	4.5142	2.0395e−05
SA−Dis [23]	36.1997	0.3669	2.9793	3.3618e−05	73.9344	0.2708	3.2971	9.7354e−05

4.3.2 定量比较

本文从每个域的测试集中随机选取了100张图像进行定量模型评估,以目标域参考图像作为风格指导来进行图像转换。表4为模型在不同数据集上的定量评价结果,每个指标最好的结果用粗体标示。

从数据集的角度来看,六个模型在CelebAHQ[17]数据集或AFHQ[17]数据集上的图像转换效果最好,对真实对象类数据集的适应性最强。对于内容特征的保持和风格特征的学习,模型在CelebAHQ[17]数据集上的完成度最高,而由于艺术类对象数据集的域间差异极大,模型对Face2Anime[54]数据集的完成度最低。观察发现,模型在Summer2Winter[4]上表现不佳,除了模型本身不适合场景类图像外,也可能因为数据集本身存在缺陷:数量小且有重复图像;一些图片中出现了大面积的人类或动物;某些图像的域特点不明显,域归属不明确。

从模型的角度来看,SAVI2I[22]的图像转换效果最好,在4个数据集的6/16项指标上取得最优表现,对不同类数据集的适应能力最强;其次是StarGAN-v2[17],在4个数据集的3/16项指标取得最优表现。虽然MUNIT[8]在CelebAHQ[17]数据集和AFHQ[17]数据集上的DIPD[13]指标最低,但从图17、图18来看,跨域转换效果不明显,甚至转换失败,导致SIFID[59]指标较高。因此,DIPD[13]指标需和SIFID[59]指标结合进行比较。

5 总结与展望

基于解耦内容-风格特征表示的图像转换模型在生成图像的质量、多样性和连续性等方面已取得了很大的进展,是图像转换模型中的重要组成部分。本文首先对图像转换进行简要介绍,梳理了基于解耦内容-风格特征表示模型的研究脉络,整理了常见数据集和评价指标,并对经典模型进行定量和定性的比较。

解耦内容和风格的I2IT模型因其"解耦"的特点在图像控制方面有着天然的优势,可以被进一步利用和挖掘。未来可探索的方向有:

(1)内容-风格特征表示的控制。现有模型缺少对内容特征和风格特征在不同类数据集上表达能力的研究,因此对不同转换任务的兼容能力不足。从控制内容和风格的角度出发,如何使模型能够根据不同数据集权衡内容和风格的保持、变化程度对构建通用转换模型有着重要意义。

(2)结构的简化。现有模型结构复杂,训练时间较长,受限于巨大的运算开销,模型通常只能对分辨率较低的图像进行处理。如何在保持模型性能的同时简化结构有待进一步探索和研究。

(3)少样本学习。基于解耦内容-风格特征表示的图像转换模型对少样本学习的研究较少,虽然目前的模型在许多大型数据集上取得了良好的效果,但应用范围受限于数据集的种类和训练时长。

参考文献(References):

[1] Isola P, Zhu J, Zhou T, et al. Image-to-image translation with conditional adversarial networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 1125-1134.

[2] Wang T, Liu M, Zhu J, et al. High-resolution image synthesis and semantic manipulation with conditional gans [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 8798-8807.

[3] Park T, Liu M, Wang T, et al. Semantic image synthesis with spatially-adaptive normalization[C]// Proceedings of the IEEE/CVF Conference on Computer Vision andPattern Recognition, 2019: 2337-2346.

[4] Zhu J, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]// Proceedings of the IEEE International Conference on Computer Vision, 2017: 2223-2232.

[5] Lee H, Tseng H, Mao Q, et al. Drit++: Diverse image-to-image translation via disentangled representations[J]. International Journal of Computer Vision, 2020, 128(10): 2402-2417.

[6] Gong R, Li W, Chen Y, et al. Dlow: Domain flow for adaptation and generalization[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 2477-2486.

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

[8] Huang X, Liu M, Belongie S, et al. Multimodal unsupervised image-to-image translation[C]//Proceedings of the European Conference on Computer Vision (ECCV), 2018: 172-189.

[9] Lee H, Tseng H, Huang J, et al. Diverse image-to-image translation via disentangled representations[C]. Proceedings of the European Conference on Computer Vision (ECCV), 2018: 35-51.

[10] Zhu JY, Zhang R, Pathak D, et al. Toward multimodal image-to-image translation[C]//Advances in Neural Information Processing Systems, 2017: 465–476.

[11] Liu MY, Breuel T, Kautz J. Unsupervised image-to-image translation networks[C]//Advances in Neural Information Processing Systems, 2017：700－708.

[12] Kim J, Kim M, Kang H, et al. U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation[DB/OL]. arXiv preprint arXiv:1907.10830, 2019.

[13] Liu M, Huang X, Mallya A, et al. Few-shot unsupervised image-to-image translation[C]//Proceedings of the IEEE/CVF international Conference on Computer Vision, 2019: 10551-10560.

[14] Yu X, Chen Y, Li T, et al. Multi-mapping image-to-image translation via learning disentanglement[C]// Advances in Neural Information Processing Systems, 2019 ：2994－3004.

[15] Wu W, Cao K, Li C, et al. Transgaga: Geometry-aware unsupervised image-to-image translation[C]//Proceedings of the IEEE/CVF Conference on Computer VisionConference on Computer Vision and Pattern Recognition, 2019: 8012-8021.

[16] Chang H, Wang Z, Chuang Y. Domain-specific mappings for generative adversarial style transfer[C]. European Conference on Computer Vision, 2020: 573-589.

[17] Choi Y, Uh Y, Yoo J, et al. Stargan v2: Diverse image synthesis for multiple domains[C]. Proceedings of the IEEE/CVF Conference on Computer VisionConference on Computer Vision and Pattern Recognition, 2020: 8188-8197.

[18] Park T, Zhu J, Wang O, et al. Swapping autoencoder for deep image manipulation[C]//Advances in Neural Information Processing Systems, 2020: 7198-7211.

[19] Jiang L, Zhang C, Huang M, et al. Tsit: A simple and versatile framework for image-to-image translation[C]// European Conference on Computer Vision, 2020: 206-222.

[20] Baek K, Choi Y, Uh Y, et al. Rethinking the truly unsupervised image-to-image translation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 14154-14163.

[21] Nederhood C, Kolkin N, Fu D, et al. Harnessing the conditioning sensorium for improved image translation[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 6752-6761.

[22] Mao Q, Tseng H, Lee H, et al. Continuous and diverse image-to-image translation via signed attribute vectors[J]. International Journal of Computer Vision, 2022, 130(2): 517-549.

[23] Kim K, Park S, Jeon E, et al. A Style-aware Discriminator for Controllable Image Translation[C]//Proceedings of the IEEE/CVF Conference on Computer VisionConference on Computer Vision and Pattern Recognition, 2022: 18239-18248.

[24] Yang S, Jiang L, Liu Z, et al. Unsupervised Image-to-Image Translation with Generative Prior[C]//Proceedings of the IEEE/CVF Conference on Computer VisionConference on Computer Vision and Pattern Recognition, 2022: 18332-18341.

[25] Kingma D P, Welling M. Auto-encoding variational bayes[DB/OL]. arXiv preprint arXiv:1312.6114, 2013.

[26] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks[J]. Communications of the ACM, 2014, 63(11): 139-144.

[27] Kim T, Cha M, Kim H, et al. Learning to discover cross-domain relations with generative adversarial networks[C]// International Conference on Machine Learning, 2017: 1857-1865.

[28] Liu Y, Sangineto E, Chen Y, et al. Smoothing the disentangled latent style space for unsupervised image-to-image translation[C]//Proceedings of the IEEE/CVF Conference on Computer VisionConference on Computer Vision and Pattern Recognition, 2021: 10785-10794.

[29] Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks[C]// Proceedings of the IEEE/CVF Conference on Computer VisionConference on Computer Vision and Pattern Recognition, 2019: 4401-4410.

[30] Ji X, Henriques J F, Vedaldi A. Invariant information clustering for unsupervised image classification and segmentation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 9865-9874.

[31] He K, Fan H, Wu Y, et al. Momentum contrast for unsupervised visual representation learning[C]// Proceedings of the IEEE/CVF Conference on Computer VisionConference on Computer Vision and Pattern Recognition, 2020: 9729-9738.

[32] Caron M, Misra I, Mairal J, et al. Unsupervised learning of visual features by contrasting cluster assignments[C]// NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems, 2020: 9912-9924.

[33] Jakab T, Gupta A, Bilen H, et al. Conditional image generation for learning the structure of visual objects[J]. methods, 2018, 43: 44.

[34] Brock A, Donahue J, Simonyan K. Large scale GAN training for high fidelity natural image synthesis[DB/OL]. arXiv preprint arXiv:1809.11096, 2018.

[35] Mao Q, Lee H, Tseng H, et al. Mode seeking generative adversarial networks for diverse image synthesis[C]//Proceedings of the IEEE/CVF Conference on Computer VisionConference on Computer Vision and Pattern Recognition, 2019: 1429-1437.

[36] Choi J, Kim D, Song B C. Style-guided and disentangled representation for robust image-to-image translation[C]// Proceedings of the AAAI Conference on Artificial Intelligence 2022,36(1).

[37] Huang X, Belongie S. Arbitrary style transfer in real-time with adaptive instance normalization[C]// Proceedings of the IEEE international Conference on Computer Vision, 2017: 1501-1510.

[38] Kwon G, Ye J C. Diagonal attention and style-based gan for content-style disentanglement in image generation and translation[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 13980-13989.

[39] Zhang R, Isola P, Efros A A, et al. The unreasonable effectiveness of deep features as a perceptual metric[C]//Proceedings of the IEEE Conference on Computer VisionConference on Computer Vision and Pattern Recognition, 2018: 586-595.

[40] Nilsback ME, Zisserman A. Automated flower classification over a large number of classes[C]// Sixth Indian Conference on Computer Vision, Graphics & Image Processing, 2008: 722-729.

[41] Krause J, Stark M, Deng J, et al. 3d object representations for fine-grained categorization[C]// Proceedings of the IEEE international Conference on Computer Vision Workshops, 2013: 554-561.

[42] Silberman N, Hoiem D, Kohli P, et al. Indoor segmentation and support inference from rgbd images[C]// European Conference on Computer Vision, 2012: 746-760.

[43] Yu F, Seff A, Zhang Y, et al. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop[DB/OL]. arXiv preprint arXiv: 1506.03365, 2015.

[44] Cordts M, Omran M, Ramos S, et al. The cityscapes dataset for semantic urban scene understanding[C]// Proceedings of the IEEE Conference on Computer VisionConference on Computer Vision and Pattern Recognition, 2016: 3213-3223.

[45] Richter S R, Vineet V, Roth S, et al. Playing for data: Ground truth from computer games[C]// European Conference on Computer Vision, 2016: 102-118.

[46] Chu W, Zheng X, Ding D. Camera as weather sensor: Estimating weather information from single images[J]. Journal of Visual Communication and Image Representation, 2017, 46: 233-249.

[47] Shen Z, Huang M, Shi J, et al. Towards instance-level image-to-image translation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 3683-3692.

[48] Sun P, Kretzschmar H, Dotiwalla X, et al. Scalability in perception for autonomous driving: Waymo open dataset[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 2446-2454.

[49] Yu F, Xian W, Chen Y, et al. Bdd100k: A diverse driving video database with scalable annotation tooling[DB/OL]. arXiv preprint arXiv:1805.04687, 2018, 2(5): 6.

[50] Zhou B, Zhao H, Puig X, et al. Semantic understanding of scenes through the ade20k dataset[J]. International Journal of Computer Vision, 2019, 127(3): 302-321.

[51] Wang X, Tang X. Face photo-sketch synthesis and recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 31(11): 1955-1967.

[52] Lee C, Liu Z, Wu L, et al. Maskgan: Towards diverse and interactive facial image manipulation[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 5549-5558.

[53] Karras T, Aittala M, Hellsten J, et al. Training generative adversarial networks with limited data[C]// NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems, 2020: 12104-12114.

[54] Li B, Zhu Y, Wang Y, et al. AniGAN: style-guided generative adversarial networks for unsupervised anime face generation[J]. IEEE Transactions on Multimedia, 2021, 24: 4077 - 4091.

[55] Heusel M, Ramsauer H, Unterthiner T, et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium[C]//NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing SystemsNeural Information Processing Systems, 2017: 6629–6640.

[56] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database[C]// IEEE Conference on Computer Vision and Pattern Recognition, 2009: 248-255.

[57] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C]// Proceedings of the IEEE Conference on Computer VisionConference on Computer Vision and Pattern Recognition, 2016: 2818-2826.

[58] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[DB/OL]. arXiv preprint arXiv:1409.1556, 2014.

[59] Shaham T R, Dekel T, Michaeli T. Singan: Learning a generative model from a single natural image[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 4570-4580.

编辑:王谦