计算机科学
COMPUTER SCIENCE

基于原型修正的小样本半监督语义图像翻译算法

何知霖, 顾天昊, 徐冠华

引用本文

何知霖, 顾天昊, 徐冠华. 基于原型修正的小样本半监督语义图像翻译算法[J]. 计算机科学, 2024, 51(8): 224-231.

HE Zhilin, GU Tianhao, XU Guanhua. Few-shot Semi-supervised Semantic Image Translation Algorithm Based on Prototype Correction [J]. Computer Science, 2024, 51(8): 224-231.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230500038

基于原型修正的小样本半监督语义图像翻译算法

何知霖[1,2]　顾天昊[1,2]　徐冠华[1]
1 青岛大学自动化学院　山东 青岛 260000
2 青岛大学智能无人系统研究院　山东 青岛 260000
(hezhilin2000@163.com)

摘　要　图像翻译任务是计算机视觉领域一个重要的研究方向,在图像风格化、超分辨率图像生成等视觉领域都有着广泛的应用。针对图像翻译任务中语义信息标注成本高、数据集通常标注困难的问题,提出了一种基于原型修正的小样本语义图像翻译算法,该算法主要包含StyleGAN、语义相似度回归器、pSp编码器模块。首先,为了降低模型对标签图像的依赖,该算法使用StyleGAN预训练模型充当生成器,增加小样本场景下的训练样本数和提升模型生成的多样性。其次,考虑到样本语义类内差异,该算法设计语义相似度回归器对原型进行修正,提升伪标签的准确率,增强模型优化效果。然后,结合标签图像和合成图像的特征图以及原型向量,实现语义信息的循环合成,构建出自监督损失函数以避免语义相似度回归器训练的标签信息需求,并利用伪标签图像对pSp编码器继续进行训练,实现语义图像翻译任务。最后,实验结果验证了所提算法在泛化性能和合成图像的多样性方面均优于经典算法。

关键词:图像翻译;原型修正;小样本学习;对抗生成网络
中图分类号　TP391

Few-shot Semi-supervised Semantic Image Translation Algorithm Based on Prototype Correction

HE Zhilin[1,2],GU Tianhao[1,2] and XU Guanhua[1]
1 School of Automation,Qingdao University,Qingdao,Shandong 260000,China
2 Institute of Intelligent Unmanned System,Qingdao University,Qingdao,Shandong 260000,China

Abstract　Image translation plays a vital role in computer vision and has extensive applications in visual fields,such as image stylization and image super-resolution generation. Datasets are frequently challenging to label,and semantic labeling has substantial costs. This paper proposes a few-shot semantic image translation framework based on prototype correction,mainly encompassing the StyleGAN module,semantic similarity regressor module,and pSp encoder module. First,to decrease the dependence of the model on the labeled image,our framework utilizes the StyleGAN pre-trained model as a generator,which expands the number of training samples in few-shot and the diversity of image generation. Second,considering the variations within the sample semantic class,our framework designs a semantic similarity regressor to correct the prototype,improving the accuracy of the pseudo-label and enhancing the model optimization effect. Third,the cyclic synthesis of semantic information is realized by combining label feature maps,synthetic feature maps and prototype vectors. Meanwhile,a self-supervised loss function is constructed to avoid the label information requirements of semantic similarity regressor training. Then the pSp encoder is trained with pseudo-tag images,and the task of semantic image synthesis is achieved. Experimental results show that the proposed framework is superior to classical frameworks in terms of excellent generalization performance and diversity of synthesized images.

Keywords　Image translation,Prototype correction,Few-shot learning,Generative adversarial network

1　引言

随着人工智能的发展,图像任务的应用日益广泛。图像翻译是其中一个重要的应用领域,它可以将一幅图像转换为另一幅图像。这项技术可应用于图像风格化转换、超分辨率图像生成、颜色填充、四季变换等视觉任务。然而,传统的图像翻译技术往往只能处理简单的语言和图像,无法很好地应对复杂的语义和视觉场景。因此,近年来语义图像翻译受到广泛关注。语义图像翻译指基于语义分割的结果来生成真实图片。完成语义图像翻译任务需提取图像样本的语义标签

到稿日期:2023-05-08　返修日期:2023-09-30
基金项目:国家自然科学基金(62076094,61773227);中国博士后科学基金(2022M721744);山东省博士后创新人才支持计划(SDBX2022023)
This work was supported by the National Natural Science Foundation of China(62076094,61773227),China Postdoctoral Science Foundation(2022M721744) and Postdoctoral Innovation Talent Support Program of Shandong Province(SDBX2022023).
通信作者:顾天昊(gutianhao@qdu.edu.cn)

进行训练，语义标签是图像生成领域最常用的条件信息，利用语义合成图像易于实现且可靠性高。

在语义图像翻译领域，条件对抗生成网络是最常见和有效的手段。Wang 等改进 pix2pix 基线算法[1]得到 pix2pixHD++算法[2]，输出具有逼真纹理的高分辨率图像，但原始的语义输入经过常规归一化层时丢掉了一些细节信息。Park 等提出了 SPADE 算法[3]，通过加入新的“空间自适应归一化”正则化层，能在不同场景中合成更逼真的图像。然而，现实生活中，语义信息标注复杂且成本高，模型训练所需数据规模不足，训练常存在过拟合的问题。因此，此类算法难以应用于现实小样本场景。为了缓解该问题，基于小样本场景的语义图像翻译成为新的研究热点并取得了一定的研究成果。Endo 等提出了使用 StyleGAN[4]先验的小样本语义图像翻译[5]算法。同时，在小样本场景下，基于 StyleGAN 模型预训练的跨域一致性技术也得到广泛研究[6-8]，但此类方法在原型构建过程中策略单一，利用标签数据回归伪标签时效率低下，合成的图像与语义信息匹配度低，结果较为粗糙。

针对上述问题，本文提出了一种基于原型修正的小样本半监督语义图像翻译（Few-shot Semi-supervised Semantic Image Translation Algorithm Based on Prototype Correction，FSISPR）算法。考虑到模型对标签图像的依赖问题，舍弃了主流框架选择的对抗生成网络中的判别器，转而选择预训练的 StyleGAN 生成器以降低训练复杂度，进一步降低对数据的依赖；同时利用 StyleGAN 生成器提取标签图像的通用表示，减少了训练生成器时要求的数据规模。此外，为了提升合成的图像与语义信息间的匹配度，设计了语义相似度回归器来估计标签图像与随机生成图像间的语义相似度，进而修正原型向量以合成伪标签。本文算法通过标签图像和合成图像各自的特征图以及原型向量实现语义信息的循环合成，进而构建出自监督损失函数以避免语义相似度回归器训练的标签信息需求。最后利用 CelebAMask 人脸数据集[9]、LSUN 数据集中的 church 图像子集[10]、FFHQ 人脸数据集[4]进行实验，通过可视化和具体的量化指标[11-12]验证了所提算法性能；同时可视化了修正原型向量的作用和效果，对该算法的创新点进行检验。

2 相关工作

2.1 语义图像翻译

近年来在图像翻译领域，对抗生成网络作为一种在数据量不足的情况下提高算法性能的重要技术，得到了广泛应用，其核心思想是通过密度匹配训练数据集密度来生成示例。当前有很多经典的结合对抗生成网络进行图像翻译的方法，这些方法通过训练生成器和鉴别器，从而实现高质量图像的合成。pix2pix 方法[1]是经典的用于图像到图像翻译的条件 GAN 框架。该方法将语义标签图和对应图像的通道级联作为判别器，生成图像的分辨率可达 512×512。但在对应用 pix2pix 框架生成的高分辨率图像进行测试时，训练不稳定，生成图像的质量不理想。因此，Wang 等在 2020 年提出了 pix2pixHD++算法[2]，该框架扩展到具有两个额外特征的交互式视觉操作，并纳入了对象实例分割信息，可以生成高达 1 024×1 024 的高分辨率图像。然而，该方法会造成语义信息损失。

为应对该问题，条件归一化层被用于控制神经网络中的特征表达和参数更新，防止信息损失。在图像翻译的相关工作中，有一些新的正则化层创新方法被提出。激活图层归一化将激活函数层与归一化层统一为一个计算图，利用层搜索算法建立了一组全新的归一化激活层 EvoNorms，实验结果表明该方法效果十分优秀[13]；注意力归一化的语义信息编辑结合了语义布局学习和局部归一化操作，以更好地增强图像翻译的语义内容的相关性[14]；为解决稀疏标签的语义信息提取问题，双重引导归一化（DGNorm）被提出，其通过加入具有大感受野的全局特征来区分同一语义区域内的激活函数，结合局部和全局特征指导图像生成[15]；自适应实例归一化（AdaIN）在进行归一化时，使用了来自另一个输入的均值和标准差来调整输入的特征图，将输入的特征映射与不同的风格信息结合，从而实现更好的图像翻译效果[16]。AdaIN 正则化层也被广泛运用于 SPADE，StyleGAN[4]和 StyleGAN2[17]等图像翻译方法中。另外，GauGAN 方法使用了 SPADE 的变体正则化层，它通过额外的条件网络来计算归一化参数，提升了性能[18]。这些创新方法不仅可以提高图像翻译的质量和效率，也为其他视觉任务提供了更好的模型设计和训练思路。但条件归一化层的性能通常依赖大量的训练数据来计算其统计量和模型参数，在小样本场景中，由于缺乏足够的训练数据，因此条件归一化层可能无法发挥其正则化的作用。

2.2 小样本

近年来，考虑现实场景需求在小样本图像生成领域已经成为新的研究热点并取得了一定的研究成果，Liu 等提出了一个小样本无监督图像翻译框架 FUNIT，实现了小样本场景下从源类别图像到目标图像的翻译[19]。但当目标类别与源图像有显著差异时，存在仅变更源图像的颜色的一些失败样本，且很难保留输入图像的结构。为了解决该问题，Saito 等提出了 COCO-FUNIT 之前的图像到图像变换模型，尝试从输入图像的不可见域条件化的示例图像中提取样式代码，并使用恒定的样式偏置设计[20]。同样，Li 等使用通道注意力双线性度量网络（CABMN）增强了模型对图片局部重要区域的关注度，并利用双线性哈达玛积操作挖掘重要区域的深层次二阶特征信息，捕捉图像的局部特征和细微差别，对保留图像结构具有重要意义[21]。此外，Ojha 等还考虑到小样本学习容易导致过拟合的问题，先用大规模的源图像域进行预训练，再将结果迁移到目标集上，此外还通过跨域距离一致性损失保留在源域中学到的物体之间的相似性和区分性[6]。但小样本学习还存在着过于依赖大规模数据集的风格内容的问题，除了获取大量图像的成本太高，某些场景也是罕见的或危险的。在这种情况下，Pizzati 等提出了 ManiFest，使用加权流形插值和局部-全局少镜头损失特征一致性，同时引入了新的残差校正机制，用于实现通用翻译和示例翻译[22]。该方法对高度非结构化的转换具有鲁棒性，包括恶劣天气生成或夜间渲染，同时也提升了性能，但对潜在空间的限制会影响该方法对图像翻译的效果，导致其难以运用于具有复杂或多样化的潜在空间的图像翻译任务中。一些研究者还将扩散模型与对抗

生成网络相结合，Wang 等以不同方式处理语义布局和噪声图像，分别输入提供给解码器和 U-Net 结构的编码器，极大地提高了语义图像合成中的生成质量和语义可解释性[23]；Careil 等则提出类亲和迁移(Class Affinity Transfer)，利用源数据集和目标数据集中类之间的相似性进行迁移学习，将源类大数据集上训练的模型应用于目标类小数据集上，以提高其学习能力[24]。

总体来说，基于小样本的图像翻译方法应用仍然有限，模型的泛化能力和稳定性仍有待提升。

3 算法

3.1 总体框架

本文提出的基于原型修正的小样本半监督语义图像翻译(FSISPR)算法主要使用了 StyleGAN[4]、pSp 编码器[25]和语义相似度回归器网络模型，结合自监督训练方式，实现图像翻译任务。该框架结构如图 1 所示。

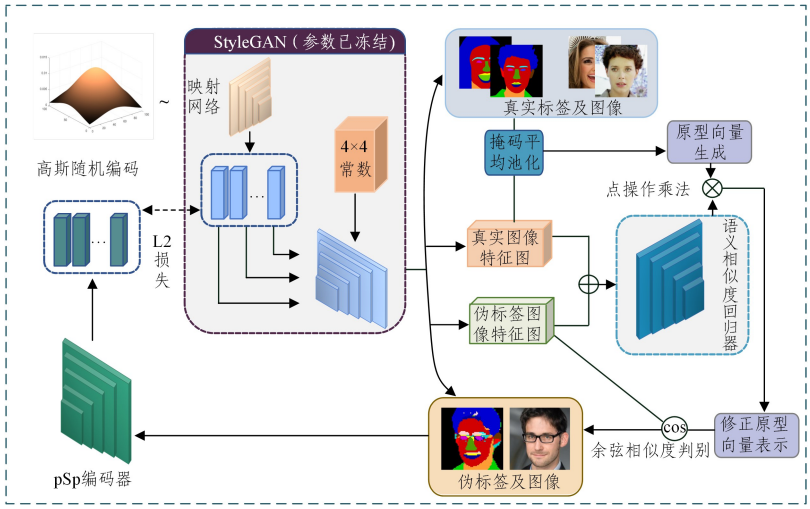


图 1 FSISPR 算法总体框架图

Fig. 1 Overall framework of FSISPR algorithm

首先，该框架借助 StyleGAN 模型[4]实现半监督学习，提取标签图像的特征图后，结合语义标签提取各语义的原型向量。其次，该框架使用语义相似度回归器对标签图像特征图与生成图像特征图的相似度进行计算，并修正提取的原型向量。接着，对修正原型向量与生成图像特征图进行逐像素比对，以余弦相似度为指标进行分类，合成伪标签，并将其作为训练数据以扩大规模。最后，生成图像和其对应的伪标签用于训练 pSp 编码器[25]，以实现语义标签到浅编码的映射，完成语义图像翻译任务。该过程中，提取原型向量和伪标签的合成是本文算法设计的关键。

本文算法修正了原型向量，增强了伪标签质量，使 pSp 编码器[25]得到了更好的训练，从而优化了整个框架的性能，详细过程将在下文中介绍。

3.2 语义相似度回归器网络模型

为了提升鲁棒性，原型向量在合成时尽可能地考虑到每个像素的贡献。但是在小样本场景中特征所能覆盖的范围有限，当同类语义的颜色、光照等场景变化时，原型向量区分各语义类别的效果明显下降。

为了使原型向量可根据伪标签图像的特征有效合成伪标签，本文算法设计了一个语义相似度回归器网络模型，如图 2 所示，以生成图像和真实图像的特征图为输入，输出生成图像与真实图像的相似度；根据图像在每个语义上的相似度对标签图像在原型向量中的占比进行调整，以修正提取的原型向量。

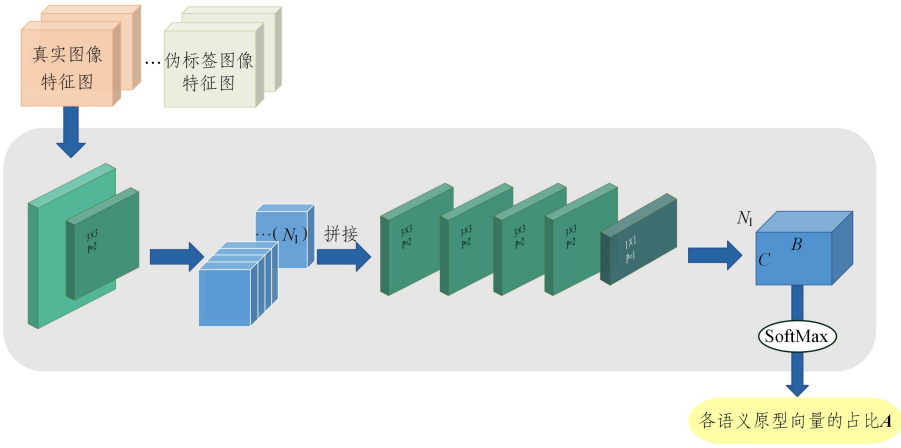


图 2 语义相似度回归器示意图

Fig. 2 Schematic diagram of semantic similarity regressor

该语义相似度回归器网络由全卷积层组成，共包含 7 个卷积层，其中前五层是步长为 2 的 3×3 卷积层，最后一层为步长为 1 的 1×1 卷积输出层。各特征图的通道数较多，故首先将它们输入第一层卷积层进行降维并进行拼接操作：

$$\boldsymbol{H}_{1,i}=conv_1(\boldsymbol{F}_i) \tag{1}$$

$$\boldsymbol{H}_1=\mathrm{cat}(\boldsymbol{H}_{1,1},\boldsymbol{H}_{1,2},\cdots,\boldsymbol{H}_{1,N_l},\boldsymbol{H}_{1,N_l+1}) \tag{2}$$

其中，$\boldsymbol{H}$ 为语义相似度回归器卷积层输出的特征图。

送入后面的卷积层挖掘相似度信息：

$$\boldsymbol{H}_6=\mathrm{conv}_{2\text{-}6}(\boldsymbol{H}_1) \tag{3}$$

接着，输出层对特征进行降维：

$$\boldsymbol{H}_7=\mathrm{reshape}(\mathrm{conv}_7(\boldsymbol{H}_6)) \tag{4}$$

其中，reshape(·)表示将特征图尺寸调整为 $B\times C\times N_l$，B 代表 batch-size 大小，C 代表语义类别数。

最后，为了得到真实图像与生成图像在不同语义上的相似度，在通道维度上进行 softmax 操作，得到各语义原型向量的占比 $\boldsymbol{A}$：

$$\boldsymbol{A}=\mathrm{softmax}(\boldsymbol{H}_7),\boldsymbol{A}\in\mathbb{R}^{B\times C\times N_l} \tag{5}$$

该占比 $\boldsymbol{A}$ 在原型向量修正操作中具有较大的作用。

3.3 FSISPR 算法

本节对 FSISPR 算法的细节进行介绍，包括原型向量提取与修正、伪标签的合成过程。

FSISPR 算法采用 one-hot 类别标注形式对每个像素进行语义分类，标签图像对 $\mathcal{D}_l$ 可表示为：

$$\mathcal{D}_l=\{\boldsymbol{x}_i,\boldsymbol{y}_i\}_{i=1}^{N_l},x\in\{0,1\}^{C\times W\times H} \tag{6}$$

其中，$\boldsymbol{x}$ 为语义标签，$W\times H=\sum_{c,x,y}^{C\times W\times H}x_{i,c,x,y}$，$W$ 和 H 分别为图像的宽和高，$\boldsymbol{y}_i$ 代表真实图像。

首先提取标签数据原型向量，其中标签图像的特征图是提取的关键。为了获得相应标签图像的特征图，本文算法尝试借助 StyleGAN 模型[4]对标签图像 y 进行反演，获得标签图映射到浅层空间的编码。该过程主要采用反推方式实现，向 StyleGAN 模型[4]输入随机噪声，生成对应图像。通过梯度向前传播和优化，该随机噪声逐渐收敛，生成与标签图像近似的图像。同时，优化后的浅层编码与标签图像形成一一对应。最后，利用这些浅层编码获取标签图像的特征图 $\boldsymbol{F}$。

接着，相应的语义标签利用迭代得到的特征图 $\boldsymbol{F}$ 对像素级特征进行分类。为使各类别的特征表示具有良好的鲁棒性，各语义的原型向量表示为该语义所有特征向量的均值[26]，表达式如下：

$$\boldsymbol{v}_c=\frac{\mathbf{1}}{N_l}\sum_{i=1}^{N_l}\frac{\sum_{x,y}\boldsymbol{F}_i^{x,y}\mathbf{1}[\boldsymbol{x}_i^{(c,x,y)}=\mathbf{1}]}{\sum_{x,y}\mathbf{1}[\boldsymbol{x}(c,x,y)_i=\mathbf{1}]} \tag{7}$$

其中，图像被看作二维平面，x 和 y 表示像素的横纵坐标；1[·]是 0-1 函数，当满足条件时输出为 1，否则为 0。

为了使原型向量可根据伪标签图像的特征有效合成伪标签，设计了语义相似度回归器以评估生成图像与各标签图像的相似度，并以此修正原型向量以及增加相似场景所占比重。因此，算法结合语义相似度回归器输出的各语义原型向量的占比 $\boldsymbol{A}$，对生成原型向量的公式进行改进，得到如下公式：

$$\hat{\boldsymbol{v}}_c=\sum_{i=1}^{N_l}\frac{a_{i,c}\sum_{x,y}\boldsymbol{F}_i^{x,y}\mathbf{1}[\boldsymbol{x}_i^{(c,x,y)}=1]}{\sum_{x,y}\mathbf{1}[\boldsymbol{x}_i^{(c,x,y)}=1]} \tag{8}$$

其中，$a_{i,c}\in\boldsymbol{A}$。

FSISPR 算法利用得到的各语义场景的相似度评估对原型向量进行修正，使得到的原型向量可根据生成图像的特征有效合成伪标签，对当前生成图像所在场景有明显倾向。

下面利用修正后的语义原型向量对生成图像的特征图上的像素级特征进行分类。分类方法采用最近邻匹配方式，将余弦相似度作为分类指标，选取余弦相似度最大的类别为该像素的语义类别，进行整体语义伪标签的合成。具体计算方式如下：

$$\boldsymbol{C}^{(x,y)}=\arg\max_{c\in C}\cos(\hat{\boldsymbol{v}}_c,\boldsymbol{F}'^{(x,y)}) \tag{9}$$

其中，$\boldsymbol{C}$ 表示像素分类结果，$\boldsymbol{F}'$ 表示生成图像特征图，cos(·)表示余弦相似度计算操作。

FSISPR 框架训练过程如算法 1 所示。

算法 1 FSISPR 框架训练过程

输入：一个标签集 $\mathcal{D}_l$

输出：浅层编码 $\hat{\omega}$

1. 使用 $\mathcal{D}_l$ 预训练 StyleGAN，计算出原型向量 $\mathbf{v}_c$ 和真实图像特征图 $\mathbf{F}$
2. for each training iteration do
3. 由 $\mathcal{N}(\mathbf{0},\mathbf{I})$ 得到样本浅层编码 $\hat{\omega}$
4. 将浅层编码反馈给生成器并得到伪标签图像特征图 $\mathbf{F}'$
5. 估计 $\mathbf{F}$ 与 $\mathbf{F}'$ 的相似度 $\mathbf{A}$
6. 修正原型向量，$\hat{\mathbf{v}}_c\leftarrow\mathbf{v}_c$
7. 反馈伪标签语义编码给 pSp 编码器
8. 计算损失函数 L_{proto} 和 L_{code}，计算公式见式(10)和式(11)
9. 计算梯度并优化框架
10. End for

3.4 损失函数

为了更好地实现语义图像翻译任务，本文的损失函数由原型一致性损失 L_{proto} 和浅层编码一致性损失 L_{code} 两部分组成。

在小样本场景下，标签图像的数量有限，为了充分训练语义相似度回归器，本文设计了一种自监督损失，即原型一致性损失，以提升合成伪标签的可靠性。根据原型一致性假设，采用余弦相似度对修正过程进行约束，即伪标签 C 结合特征图 $\boldsymbol{F}'$ 生成的伪原型向量 $\boldsymbol{v}_{u,c}$ 与已修正的原型向量 $\hat{\boldsymbol{v}}_c$，数学式表示如下：

$$L_{\text{proto}}=1-\cos(\hat{\boldsymbol{v}}_c,\boldsymbol{v}_{u,c}) \tag{10}$$

由于 pSp 编码器的有效性直接决定任务的性能，故利用浅层编码一致性损失训练 pSp 编码器，将合成的伪标签回归出浅编码 $\hat{\omega}$，与 ω 的对比形成损失函数：

$$L_{\text{code}}=|\hat{\omega}-\omega| \tag{11}$$

4 实验

4.1 数据集

实验共选取 3 个公开数据集进行对比和分析：CelebAMask 人脸数据集[9]、LSUN 数据集中的 church 子集[10]、FFHQ 数据集[4]。后两个数据集用于 StyleGAN 模型[4]的预训练，以对应本框架后续训练所需的随机图像翻译。

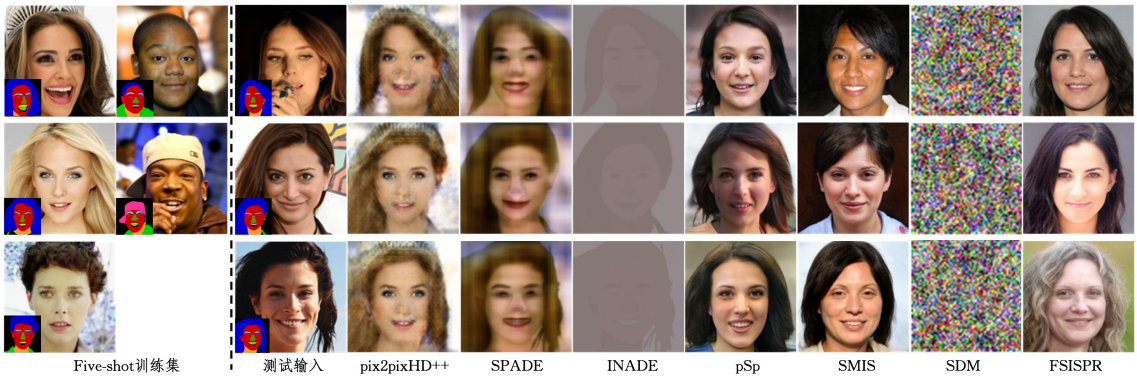
4.2 实验细节

实验采用 Pytorch 框架，在 RTX2080Ti 计算平台上实现。为了合理分析算法性能的优越性，实验共选择 6 种对比算法：pix2pixHD ++[2]，SPADE[3]，INADE[27]，pSp[25]，SMIS[28]和 SDM[22]。其中，pix2pixHD++[2]和 SPADE[3]是近年来语义图像翻译领域的顶级算法，INADE[27]是该领域生成图像多样性高的代表性算法，pSp[25]和 SMIS[28]是当前性能最突出的算法，SDM[23]是结合扩散模型与对抗生成网络的进行语义图像合成的最新方法。这 6 种算法和 FSISPR 算法将利用上述数据集进行性能比较。需要说明的是，尽管各数据集可获得大规模的训练数据，但为了在小样本场景展开研究，实验采用 five-shot 和 one-shot 学习方式对框架进行训练。在训练中，本框架将学习率定为 0.0001 并选择 Ranger 优化器进行优化学习，batch-size 以及训练最大次数分别被设置为 2 和 100000，其余各算法的参数均参考其对应文献中的定义，保证一致性。

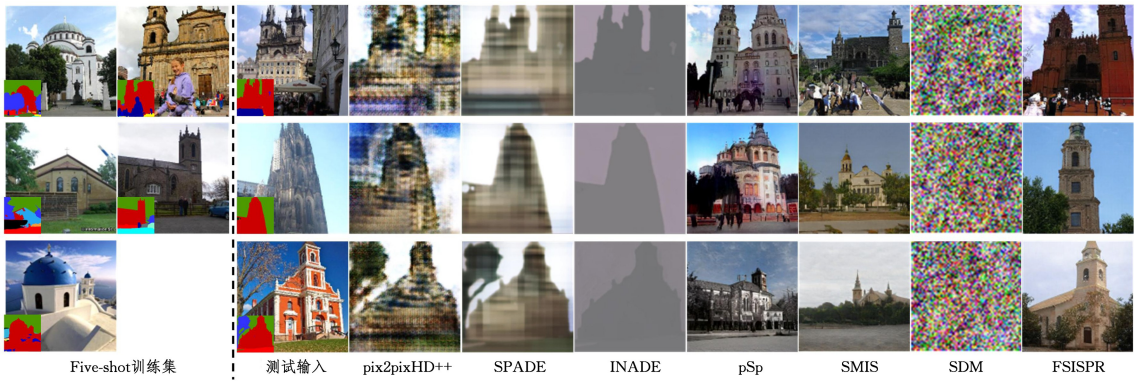
4.3 可视化结果与分析

在图像翻译任务中，合成图像的可视化效果是性能差异最直观的体现，可准确地反映算法优劣。实验分别在 five-shot 和 one-shot 学习模式下进行训练并测试。

首先对 five-shot 模式进行性能评估。图像翻译任务难度会随着语义类别数量的增加而增大，使各语义对应关系更难获得。本实验除了使用 CelebAMask 数据集，还选取了语义类别数较多的 Lsunchurch 数据集进行算法性能验证，以保证算法评估的准确性和全面性。两个数据集上的训练、测试图像以及各算法测试结果如图 3 所示。



(a) CelebAMask



(b) Lsunchurch

图 3 Five-shot 学习模式的 CelebAMask 数据集、Lsunchurch 数据集图像翻译测试结果

Fig. 3 Image translation test results of CelebAMask and Lsunchurch datasets in five-shot learning mode

由图 3(a)可知，FSISPR 框架性能明显优于其他对比算法，图像生成质量和语义对齐程度均有一定提升。pix2pixHD++算法[2]和 SPADE 算法[3]虽然语义对齐精度高，但这两者生成的图像质量较差，而 INADE 算法[27]在小样本场景下语义对齐度和图像质量都较差。相比之下，pSp 算法[25]在完成任务的过程中借助预训练的 StyleGAN 模型[4]，可有效地生成高质量人脸图像，但 pSp 编码器仅靠 5 张图像无法可靠地学习到语义到浅编码的映射，其合成的人脸相似，出现过拟合现象。与 FSISPR 框架类似，SMIS 算法[28]借助预训练的 StyleGAN 合成了大量伪标签图像对，扩充了训练数据的规模并有效提升了算法的性能，但该算法提取的原型向量差，这相当于在训练过程中引入了噪声，导致 SMIS 算法[28]合成的图像粗糙。SDM 算法[23]在小样本场景下通过扩散模型也难以学习到有效的信息，加之引入了大量的噪声，导致在其反演机制下无法有效地合成可识别的图像。本文设计的原型修正方式有效地调节原型向量形成时的线性组合，使其更倾向于语义相似度更高的图像原型，从而提升了伪标签的准确性。从图 3(a)也可看出，FSISPR 框架测试结果语义对齐程度有明显改善。如图 3(b)所示，各算法在 Lsunchurch 数据集上的图像翻译性能均存在不同程度的下降。但相比之下，FSISPR 框架依然获得了最优性能，大致恢复出了对应语义的物体。其中，pix2pixHD++算法[2]和 SPADE 算法[3]仅能大致合成各物体的形状，无法细分出不同语义，图像质量差；INADE 算法[27]仅靠少量图像训练无法提取到准确

的语义信息，导致图像质量差；pSp 算法[25]合成了质量良好的图像，但语义匹配度较低，生成图像的特征过于相似；性能较为出色的 SMIS 算法[28]则性能下降较多，因为该算法语义类别数量过多，但可用信息较少，以至于一些类别的原型向量十分相似；与合成人脸相似，在合成建筑时，SDM 算法[23]仍然存在噪声过大、合成图像质量差的问题。FSISPR 框架在语义类别较大时，由于伪标签的合成引入了大量的噪声，加上语义相似度回归器的训练难度增大，训练效果下降，训练集扩充效果不佳，但语义相似度回归器仍在一定程度上有助于提升伪标签合成的可靠性，进而帮助框架完成语义图像翻译，因此 FSISPR 框架的性能依然较为出色。

相较于 five-shot 学习模式，各模型在 one-shot 学习模式中能提取到的有效语义信息则更少，甚至缺乏部分语义信息的表达。所以该模式作为典型的小样本学习场景，可测试出极端条件下模型的性能。因此，本文选取 CelebAMask 人脸数据集进行 one-shot 学习模型训练和测试，以借助准确标签实现精确评估。由于 SDM 算法[23]在 five-shot 上表现极差，因此这里不再进行 one-shot 实验对比，其余 5 种算法的实验结果如图 4 所示。

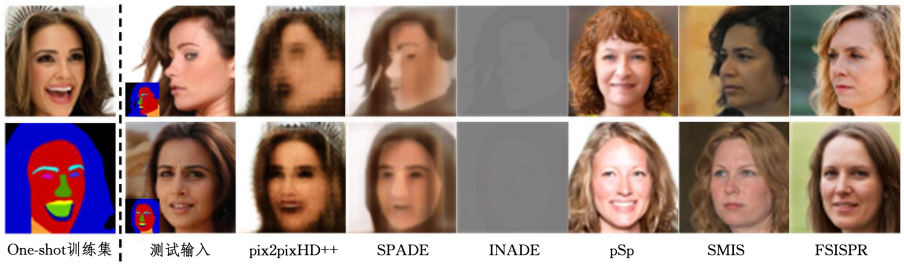


图 4 One-shot 模式下 CelebAMask 数据集图像翻译结果

Fig. 4 Image translation results of CelebAMask dataset in one-shot learning mode

由图 4 可观察到，FSISPR 框架依旧能够有效地完成训练并获得可靠性能。pix2pixHD＋＋算法[2]在合成人脸头发部分时，均带有标签图像中的头饰信息。这表明该模型在学习过程中将标签图像特有的信息当作了通用特征，因此合成的结果较为单一，多样性不足。SPADE 算法[3]和 pSp 算法[25]也在很大程度上受到该问题的影响，两者合成结果的同语义区域特点类似，多样性不足。INADE 算法[27]仅能区分人脸五官边缘和颜色深浅，无法捕捉到更多细节，说明该模型在极端小样本场景下性能较差。不同于前几种算法，SMIS 算法[28]和 FSISPR 算法借助 StyleGAN 预训练模型逐步实现了图像规模扩充，模型学习到了更多的可用特征，弥补了特征不足的问题。相比之下，FSISPR 算法经过修正后的原型向量与各场景匹配度更高，因此在合成细节上取得了比 SMIS 算法[28]更好的效果，例如人脸图像中头发的区域匹配度更高，鼻孔的大小和角度也十分吻合。

总体来说，FSISPR 框架在 five-shot 和 one-shot 模式下的生成结果都优于其他算法，但当语义类别增多时，合成伪标签引入的大量噪声影响了训练结果的可靠性。如何避免噪声的影响是未来值得研究的工作。在 one-shot 学习模式上，本文算法的原型向量经过修正，得到的结果与场景匹配度更高，合成细节更优。但是本文方法在语义信息有效性低和场景标注准确率低的场景下，存在图像语义匹配度不足的情况。

4.4 数值结果与分析

除了可视化合成图像以进行直观对比外，量化指标也是进行算法性能对比的重要方式。由于 CelebAMask 人脸数据集包含两种形式的语义标签且标签准确可靠，因此实验选择 CelebAMask 人脸数据集进行量化指标计算。同时，选取了 3 种不同的指标分别在两种标签形式上对算法性能进行评估：平均交并比(mIoU)[11]、像素精度(PA)[11]、弗氏分布距离(FID)[12]。其中 pix2pixHD＋＋[2]和 SPADE[3]两种算法在 one-shot 学习模式下效果较差，故实验未计算该模式下的指标，但补充了它们在图像扩充场景下的结果；SDM 算法[23]在小样本场景的可视化结果太差，这里不再做数值结果对比。各算法在小样本场景中的实验指标计算结果如表 1 所列。

表 1 小样本学习机制下的语义合成图像量化结果

Table 1 Quantitative results of semantic synthesis images in few-shot learning mode

Methods	N_l	CelebAMask		
		mIoU ↑	PA ↑	FID ↓
pix2pixHD＋＋[2]	5	62.2	92.7	82.7
pix2pixHD＋＋*[2]	5	38.7	87.6	98.0
SPADE[3]	5	64.8	92.4	79.3
SPADE*[3]	5	40.1	88.8	88.5
INADE[27]	1	52.8	69.8	119.8
INADE[27]	5	57.3	72.6	97.8
pSp[24]	1	24.8	61.7	93.4
pSp[24]	5	28.4	67.6	96.9
SMIS[28]	1	38.3	81.7	59.1
SMIS[28]	5	41.5	82.5	53.9
FSISPR	1	40.2	83.5	54.3
FSISPR	5	43.6	85.9	48.7

注：* 代表应用于图像扩充场景的算法性能。

结合 3 种量化指标的结果可得：对比合成的图像与真实图像的分布相似度，FSISPR 算法最高，合成图像的质量也最高；对比语义匹配度的性能，相较于 SPADE 算法[3]和 pix2pixHD＋＋算法[2]，FSISPR 算法的语义匹配效果较差。但在所有使用了图像规模扩充技术的框架中，FSISPR 框架性能最佳，这也揭示了该框架原型修正的有效性。

具体而言，pix2pixHD＋＋算法[2]和 SPADE 算法[3]的语义匹配度最高，但是它们合成的图像过于模糊，与真实图像相差较大，无法有效完成任务。在进行图像扩充后，这两种算法并未获得图像质量上的提升，语义匹配度反而出现了下降，这说明伪标签的生成包含大量噪声，以致于影响了训练效果。同样，INADE 算法[27]和 pSp 算法[25]所合成的图像与语义标签差距较大，而且 pSp 算法[25]的合成图往往过拟合，与标签

图像过于相似，导致其分布无法匹配真实图像的分布。在该问题上，SMIS 算法[28]借助伪标签的合成有效地扩充了训练集，获取了更多可用信息，从而提升了合成图像的质量，FID 指标明显下降。但是 SMIS 算法合成伪标签时准确率低，造成了合成图像的语义匹配度过于粗糙。此外，FSISPR 算法在 one-shot 学习模式下取得的语义相似度量化指标优于 SMIS 算法[28]在 one-shot 学习模式中的结果，这突显出 FSISPR 框架原型修正带来的模型鲁棒性，降低了对场景的依赖性。

总体来说，FSISPR 算法利用原型修正算法提升了伪标签的效果，进而使得语义匹配度得到提升，同时保证了图像分布的准确性。实验结果也证实 FSISPR 算法有效地提升了语义匹配指标。

4.5 原型向量修正性能评估

本文还通过实验验证了所提出的原型向量修正算法的有效性。在该实验中，通过记录多个数据集在训练过程中伪标签合成的变化，以展示修正的效果。如图 5 所示，在 CelebA-Mask 数据集上进行了训练，训练过程中，定义每 6000 次迭代为一个间隔，并展示了训练集和训练过程中的中间结果。本文的原始伪标签合成方法与 SMIS 算法相同。从训练结果可以看出，本文提出的算法的语义图像翻译效果逐渐逼近于随机采样合成的图像。当训练迭代次数达到 36000 次时，两幅图像已经十分接近，这证明了扩充图像规模的有效性。仔细观察训练中伪标签的细节可以发现，原始的伪标签合成受到标签图像场景控制，特别是当标签图像同语义区域场景多有不同时，所得到的原型向量很难准确地对特征进行分类，这给编码器的训练带来了很大的影响。为了解决该问题，本文引入了语义相似度回归器。从训练过程中可以看出，在最开始时，语义相似度回归器没有得到训练，两组伪标签几乎是一致的。但是随着训练的进行，语义相似度回归器逐渐起到了调整原型构成的作用，两组伪标签逐渐区分开来，并且修正后的伪标签逐渐得到了优化，准确率提升。例如，在修正之前，额头和头发部分的伪标签很难准确判断分类，但是在修正之后，各部分语义的分类效果都在一定程度上获得了提升。

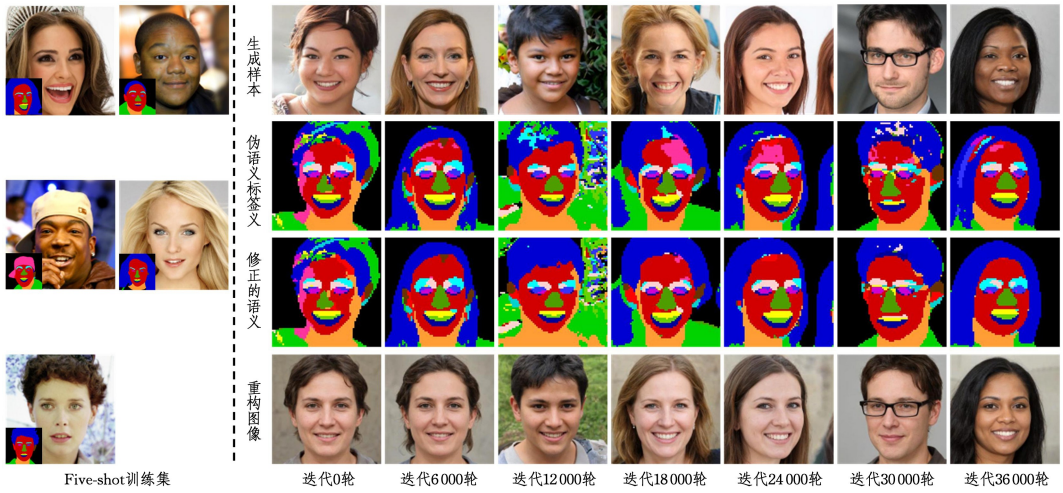


图 5 five-shot 学习模式下 CelebAMask 数据集原型修正性能可视化结果

Fig. 5 Visualization results of prototype correction performance on CelebAMask dataset in five-shot learning mode

结束语 针对小样本场景中语义图像翻译任务所存在的问题与不足，本文借鉴半监督学习的思想，设计了一种基于原型修正的小样本语义图像翻译(FSISPR)算法。该算法首先借助 StyleGAN 预训练模型获取标签图像的语义原型向量以及特征图；其次引入新型的语义相似度回归器对比标签图像与随机合成图像的语义相似度，并利用该相似度对原型进行修正，获取更准确的通用表示；然后通过已修正的原型向量对随机合成图像进行语义标注，合成伪标签用于训练集扩充；最后利用伪标签图像对编码器进行继续训练，并实现语义图像翻译任务。相较于当前已有方法，本文算法修正了原型向量并提升了伪标签的质量，使得半监督的训练可以更好地完成。

但是本文算法仍有一些不足。首先，图像语义匹配度依旧不足，其原因在于图像规模提供的有效信息少，一些场景缺乏准确标注，甚至出现部分语义信息表达缺失的情况。其次，本文的语义合成网络可以进一步调整以提高鲁棒性。未来将针对以上问题继续展开研究。

参 考 文 献

[1] LAWRENCE N, JORDAN M. Semi-supervised learning via Gaussian processes [J]. Advances in Neural Information Processing Systems, 2004, 17: 753-760.

[2] WANG T C, LIU M Y, ZHU J Y, et al. High-resolution image synthesis and semantic manipulation with conditional gans[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2018: 8798-8807.

[3] PARK T, LIU M Y, WANG T C, et al. Semantic image synthesis with spatially-adaptive normalization[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2019: 2337-2346.

[4] KARRAS T, LAINE S, AILA T. A style-based generator architecture for generative adversarial networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York: IEEE Press, 2019: 4401-4410.

[5] ENDO Y, KANAMORI Y. Few-shot semantic image synthesis

using stylegan prior [J]. arXiv:2103.14877,2021.

[6] OJHA U,LI Y,LU J,et al. Few-shot image generation via cross-domain correspondence[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York:IEEE Press,2021:10743-10752.

[7] ROBB E,CHU W S,KUMAR A,et al. Few-shot Adaptation of Generative Adversarial Net-works[J]. arXiv:2010.11943,2020.

[8] BAZAZIAN D,CALWAY A,DAMEN D. Dual-Domain Image Synthesis using Segmenta-tion-Guided GAN[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York:IEEE Press,2022:506-515.

[9] LEE C H,LIU Z,WU L,et al. Maskgan:Towards diverse and interactive facial image manipulation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York:IEEE Press,2020:5549-5558.

[10] YU F,SEFF A,ZHANG Y,et al. Lsun:Construction of a large-scale image dataset using deep learning with humans in the loop [J]. arXiv:1506.03365,2015.

[11] GARCIA-GARCIA A,ORTS-ESCOLANO S,OPREA S,et al. A review on deep learning techniques applied to semantic segmentation [J]. arXiv:1704.06857,2017.

[12] HEUSEL M,RAMSAUER H,UNTERTHINER T,et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium [J]. Advances in Neural Information Processing Systems,2017,30:6626-6637.

[13] LIU H,BROCK A,SIMONYAN K,et al. Evolving normalization-activation layers [J]. Advances in Neural Information Processing Systems,2020,33:13539-13550.

[14] WANG Y,CHEN Y C,ZHANG X,et al. Attentive normalization for conditional image generation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York:IEEE Press,2020:5094-5103.

[15] ZHU J C,GAO L L,SONG J K,et al. Label-Guided Generative Adversarial Network for Realistic Image Synthesis [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2023,45(3):3311-3328.

[16] HUANG X,BELONGIE S. Arbitrary style transfer in real-time with adaptive instance normalization[C]// Proceedings of the IEEE International Conference on Computer Vision. New York:IEEE Press,2017:1501-1510.

[17] KARRAS T,LAINE S,AITTALA M,et al. Analyzing and improving the image quality of stylegan[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York:IEEE Press,2020:8110-8119.

[18] PARK T,LIU M Y,WANG T C,et al. Gaugan:semantic image synthesis with spatially adaptive normalization[C]// International Conference on Computer Graphics and Interactive Techniques. ACM,2019.

[19] LIU M Y,HUANG X,MALLYA A,et al. Few-shot unsupervised image-to-image translation[C]// Proceedings of the IEEE International Conference on Computer Vision. New York:IEEE Press,2019:10551-10560.

[20] SAITO K,SAENKO K,LIU M Y. Coco-funit:Few-shot unsupervised image translation with a content conditioned style encoder[C]// Proceedings of Computer Vision-ECCV 2020. Berlin:Springer,2020:382-398.

[21] LI X X,AN W J,WU J J,et al. Channel attention bilinear metric network[J]. Journal of Jilin University(Engineering and Technology Edition),2024,54(2):524-532.

[22] PIZZATI F,LALONDE J F,DE. CHARETTE R. Manifest:Man-ifold deformation for few-shot image transla-tion[C]// Proceedings of Computer Vision(ECCV 2022). Berlin:Springer,2022:440-456.

[23] WANG W,BAO J,ZHOU W,et al. Semantic image synthesis via diffusion models [J]. arXiv:2207.00050,2022.

[24] CAREIL M,VERBEEK J,LATHUILIÈRE S. Few-shot Semantic Image Synthesis with Class Af-finity Transfer[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York:IEEE Press,2023:23611-23620.

[25] RICHAEDSON E,ALALUF Y,PATASHNIK O,et al. Encoding in style:a stylegan encoder for image-to-image translation [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York:IEEE Press,2021:2287-2296.

[26] WANG Y,KHAN S,GONZALEZ-GARCIA A,et al. Semi-supervised learning for few-shot image-to-image translation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York:IEEE Press,2020,4453-4462.

[27] TAN Z,CHAI M,CHEN D,et al. Diverse semantic image synthesis via probability distribution modeling[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York:IEEE Press,2021:7962-7971.

[28] ZHU Z,XU Z,YOU A,et al. Semantically Multi-Modal Image Synthesis[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. New York:IEEE Press,2020,5466-5475.

HE Zhilin, born in 2000, postgraduate, is a member of CCF(No. P8040G). Her main research interests include machine learning and visual navigation.

GU Tianhao, born in 1990, Ph.D, postgraduate supervisor. His main research interests include pattern recognition, machine learning, visual navigation and deep space exploration.

（责任编辑：何杨）