

## 融合语义标签和噪声先验的图像生成

张素素,倪建成\*,周子力,侯杰

(曲阜师范大学软件学院,山东曲阜 273165)

(\*通信作者电子邮箱nijch@163.com)

**摘要:**针对现有生成模型难以直接从复杂语义标签生成高分辨率图像的问题,提出了融合语义标签和噪声先验的生成对抗网络(SLNP-GAN)。首先,直接输入语义标签(包含形状、位置和类别等信息),使用全局生成器对其进行编码,并结合噪声先验来学习粗粒度的全局属性,初步合成低分辨率图像;然后,基于注意力机制,使用局部细化生成器来查询低分辨率图像子区域对应的高分辨率子标签,获取细粒度信息,从而生成纹理清晰的复杂图像;最后,采用改进的引入动量的Adam算法(AMM)算法来优化对抗训练。实验结果表明,与现有方法text2img相比,所提方法的像素精确度(PA)在COCO\_Stuff和ADE20K数据集上分别提高了23.73%和11.09%;相较于Adam算法,AMM算法收敛速度提升了约一倍,且损失值波幅较小。可见,SLNP-GAN能高效地获取全局特征和局部纹理,生成细粒度、高质量的图像。

**关键词:**语义标签;噪声先验;注意力机制;引入动量的Adam算法;生成对抗网络

**中图分类号:**TP391.4 **文献标志码:**A

### Image generation based on semantic labels and noise prior

ZHANG Susu, NI Jiancheng\*, ZHOU Zili, HOU Jie

(School of Software, Qufu Normal University, Qufu Shandong 273165, China)

**Abstract:** Existing generation models have difficulty in directly generating high-resolution images from complex semantic labels. Thus, a Generative Adversarial Network based on Semantic Labels and Noise Prior (SLNP-GAN) was proposed. Firstly, the semantic labels (including information of shape, position and category) were directly used as input, the global generator was used to encode them, the coarse-grained global attributes were learned by combining the noise prior, and the low-resolution images were generated. Then, with the attention mechanism, the local refined generator was used to query the high-resolution sub-labels corresponding to the sub-regions of the low-resolution images, and the fine-grained information was obtained, the complex images with clear textures were thus generated. Finally, the improved Adam with Momentum (AMM) algorithm was introduced to optimize the adversarial training. The experimental results show that, compared with the existing method text2img, the proposed method has the Pixel Accuracy (PA) increased by 23.73% and 11.09% respectively on COCO\_Stuff and the ADE20K datasets; in comparison with the Adam algorithm, the AMM algorithm doubles the convergence speed with much smaller loss amplitude. It proves that SLNP-GAN can efficiently obtain global features as well as local textures and generate fine-grained high-quality images.

**Key words:** semantic label; noise prior; attention mechanism; Adam with Momentum (AMM) algorithm; Generative Adversarial Network (GAN)

## 0 引言

图像生成任务近年来已成为计算机视觉领域的研究重点,传统的拍摄技术易受到时空的限制,无法凭空产生不存在的事物。深度学习中的图像生成技术不仅能自动为艺术家和用户生成图像,有助于视觉理解,而且还推动了跨视觉-语言推理的研究<sup>[1]</sup>,因此,准确有效地生成高分辨率图像成为目前研究的关键问题之一。

传统的图像生成主要采用非参数化生成模型和参数化生成模型:非参数化生成模型的基本思想是从数据库中匹配图

像块,主要应用于图片纹理合成和半自动图像修复;参数化的图像生成技术中的自回归方法<sup>[2]</sup>调节所有先前像素上的每个像素为概率似然建模。由于传统模型直接用数据样本进行参数更新,公式推导较繁杂且模型计算量较大。

近年来,深度学习在图像生成领域取得了较好的成果,变分自编码器(Variational Auto-Encoder, VAE)<sup>[3]</sup>使用变分推理联合学习编码器和解码器到隐码和图像的映射。随后,级联优化网络(Cascaded Refinement Network, CRN)<sup>[4]</sup>使用多个分辨率倍增的模块,从真实语义分割图生成街景的高分辨率图像。图像-图像翻译模型<sup>[5]</sup>进一步使用输入-输出图像对作为

**收稿日期:**2019-10-16; **修回日期:**2019-12-11; **录用日期:**2019-12-17。 **基金项目:**国家自然科学基金青年科学基金资助项目(61601261);山东省研究生教育教育质量提升计划项目(SDY17136);曲阜师范大学交叉学科研究项目(QFNUSKC291809120)。

**作者简介:**张素素(1997—),女,山东菏泽人,硕士研究生,主要研究方向:计算机视觉、图像生成、深度学习;倪建成(1971—),男,山东济宁人,教授,博士,CCF高级会员,主要研究方向:计算机视觉、机器学习、分布式计算;周子力(1973—),男,山东菏泽人,副教授,博士,CCF高级会员,主要研究方向:智能信息处理、嵌入式智能、知识工程;侯杰(1996—),女,山东济宁人,硕士研究生,主要研究方向:计算机视觉、深度学习。

训练数据,将输入图像转换为另一个图像域。目前,生成对抗网络(Generative Adversarial Network, GAN)是最常用的生成模型,其联合学习生成器和判别器。基于GAN的图像生成通常以文本为输入,已在简单数据集上(如鸟、花和人脸)生成了逼真的个体图像,但在包含多个对象和场景信息的数据集上难以生成高质量的复杂图像。将文本作为输入,仅以全局句子向量为条件,在单个实例级别上错过了相关信息,难以生成高质量的复杂图像<sup>[6]</sup>,如Hong等<sup>[7]</sup>提出文本到图像的生成方法(text2img),由于简短文本描述中的模糊性,对象的位置和大小未知,使生成过程难以约束。相较于文本结构,Johnson等<sup>[8]</sup>提出从场景图到图像生成方法(sg2im),由于场景图是较清晰的结构化表示,可用于编码对象、属性和关系,克服了文本输入的模糊性;但是场景图缺乏属性与空间信息,生成的图像分辨率较低且纹理较为模糊。

此外,噪声作为GAN输入的重要部分,包含了许多图像特征信息,但现有方法仅输入随机噪声,无法学习到图像属性信息<sup>[9]</sup>;同时,用引入动量的Adam(Adam with Momentum, AMM)算法<sup>[10]</sup>优化对抗训练,可解决Adam算法出现的模式崩溃和收敛速度慢等问题。

针对以上挑战和限制,本文使用基于语义标签和噪声先验的生成对抗网络SLNP-GAN进行图像生成。1)为克服文本描述的模糊性和场景结构的复杂性,直接使用语义标签作为输入,其包含了对象位置、空间关系、大小、形状等信息;2)为使得图像生成器学习到实例的全局属性并使得生成图像匹配输入的语义标签,采用有先验知识的噪声快速搜索到图像的特征,初步生成图像,再结合注意力机制合成高分辨率图像;3)为优化训练过程,使其更稳定且收敛更快,用AMM优化算法代替常用的Adam算法,提高图像生成的效率。

### 1 SLNP-GAN 模型

GAN是常用的图像生成模型<sup>[11]</sup>,包括:生成器(G)和判别器(D)。生成器主要用于学习真实图像的像素分布,使自身生成的图像更加真实;判别器需要区分接收的图像真假。生成器和判别器进行最小最大值的训练,两个模型相对抗最后达到全局最优。AttnGAN<sup>[12]</sup>将注意力机制<sup>[13]</sup>引入到图像生成中,但该模型的输入仍是简单文本形式,传递的信息有限且缺乏核心的空间属性规范,难以生成有复杂位置关系的高质量图像。图像-图像转换pix2pixHD<sup>[14]</sup>,由语义标签对应的语义布局生成了具有多个实例-关系复杂图像;然而,由语义布局生成图像是一对多问题,许多图像可能布局一致,不同外观的对象布局也可能相同,使用全局对象特征,图像缺少纹理细节,丢失了实例级别的细粒度信息。

因此,为了生成高分辨率且匹配输入语义标签类别和布局的图像,提出了融合语义标签和噪声先验的图像生成模型SLNP-GAN,模型概述如图1所示。该模型首先直接用语义标签作为输入,同时结合噪声先验初步生成低分辨率的全局图像;再使用注意力机制引导局部细化生成器进行像素级的合成,进一步生成高分辨率的图像。即:1)输入语义标签 $L_0$ ,其提供了对象类别、位置大小和形状等信息,全局图像生成器 $G_{img_0}$ 利用 $L_0$ 全局嵌入向量和噪声先验初步生成全局图像 $I_0$ ;

2)由局部细化生成器 $G_{img_1}$ 结合注意力机制,为生成的低分辨率图像的每个实例查询对应于语义标签 $L_1$ 的类标签,细化不同的区域,生成更高分辨率的图像 $I_1$ 。同时使用相同结构的多级判别器判别生成图像的真假,改变输入的语义标签可实现不同图像的生成。

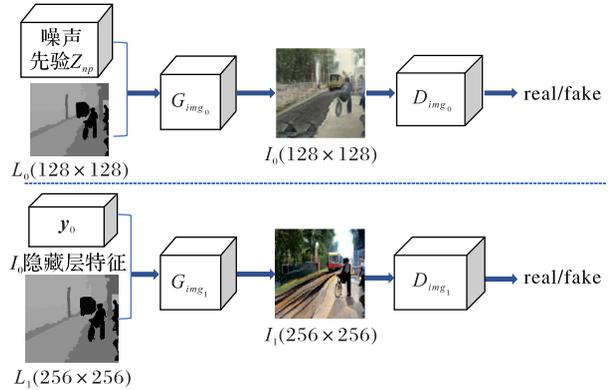


图1 SLNP-GAN模型概述

Fig. 1 Overview of SLNP-GAN model

#### 1.1 噪声先验生成机制

早期研究发现噪声分布实际代表图像属性和类别等特征信息。作为生成对抗网络的输入,若先让噪声学习到图像特征再输入网络,则生成图像的准确度能得到提升<sup>[15]</sup>。

噪声先验知识的学习方法是基于变分自编码器,其由编码器和解码器构成。变分自编码器的训练过程,如图2所示。编码器对输入高维数据进行编码得到低维隐藏层的表达,解码器对低维隐藏层解码来重构和输入大小相同的高维输出,输入和输出之间的重构误差则是模型优化的目标函数。用 $x$ 表示输入图像, $h$ 表示潜在变量, $\hat{x}$ 是重构图像。理想情况下,训练输出的重构图像应该和原图相似。编码器(Enc)和解码器(Dec)分别如式(1)、(2)表示:

$$h \sim \text{Enc}(x) = q_\theta(h|x) \tag{1}$$

$$\hat{x} \sim \text{Dec}(h) = p_\phi(x|h) \tag{2}$$

其中:分布 $q$ 和分布 $p$ 分别被 $\theta$ 和 $\Phi$ 参数化,使网络从 $x$ 映射为潜在特征向量 $h$ ,并由 $h$ 重构图像 $x$ ,最后还原为输出图像 $\hat{x}$ 。

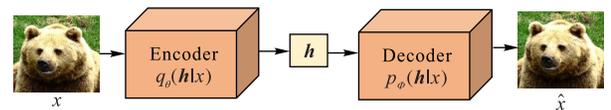


图2 变分自编码器训练过程

Fig. 2 Training process of variational autoencoder

因此,在VAE模型的基础上改进,使用先验知识的学习方法,尝试在投入模型之前先让噪声习得图像的实例属性,如图3所示。

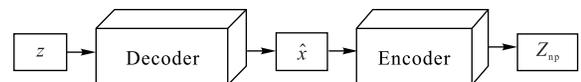


图3 噪声先验生成机制

Fig. 3 Noise prior generation mechanism

与VAE模型不同,本文先将随机噪声 $z$ 输入到在数据集上训练好的VAE解码器中生成图像 $\hat{x}$ ,再将其作为VAE编码器输入,得到包含图像属性的噪声先验 $Z_{np}$ ,作为模型的输入以生成图像。噪声先验的具体生成过程,如图4所示。

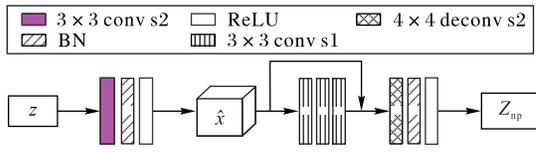


图 4 噪声先验生成过程

Fig. 4 Process of noise prior generation

噪声先验生成机制采用解码-编码的架构。首先对随机噪声下采样,该下采样模块由步长为 2 的  $3 \times 3$  卷积层、批量归一化层和 ReLU 激活层构成,通过对随机噪声解码获得图像  $\hat{x}$ ;然后将初步获得的图像特征喂入由 3 个步长为 1 的  $3 \times 3$  卷积层和一个残差连接构成的残差单元,该残差模块使得网络有更深的编码结构;最后,对获取的图像特征上采样进行编码以获得噪声先验,上采样模块由步长为 2 的  $4 \times 4$  反卷积层、批量归一化和 ReLU 激活层构成。解码器、编码器分别如式(3)、(4)所示:

$$\hat{x} \sim \text{Dec}(z) = p(\hat{x}|z) \quad (3)$$

$$Z_{np} \sim \text{Enc}(\hat{x}) = q(Z_{np}|\hat{x}) \quad (4)$$

训练随机噪声获得噪声先验,目标损失函数被定义为对数似然和先验正则项之和:

$$\mathcal{L} = \mathbb{E}_{p(\hat{x}|z)} \left[ \log \frac{q(Z_{np}|\hat{x})q(\hat{x})}{p(\hat{x}|z)} \right] = \mathcal{L}_{\text{like}}^{\text{pixel}} + \mathcal{L}_{\text{prior}} \quad (5)$$

$$\mathcal{L}_{\text{like}}^{\text{pixel}} = -\mathbb{E}_{p(\hat{x}|z)} [\log q(Z_{np}|\hat{x})] \quad (6)$$

$$\mathcal{L}_{\text{prior}} = D_{\text{KL}}(p(\hat{x}|z) \parallel q(\hat{x})) \quad (7)$$

其中:  $z \sim N(0, 1)$  是服从正态分布的随机噪声,  $D_{\text{KL}}$  是 KL 散度,  $\mathcal{L}_{\text{like}}^{\text{pixel}}$  使用对数似然表示重构误差,  $\mathcal{L}_{\text{prior}}$  表示先验正则项。加入该正则项以防止出现过拟合,同时确保模型重构的噪声先验尽可能准确。训练过程中  $p(\hat{x}|z)$  和  $q(\hat{x})$  应尽可能接近,以

最小化 KL 散度。

由此,通过对随机噪声进行预训练,对潜在分布  $q(\hat{x})$  增加先验约束,舍弃了与现实相违背的噪声数据,生成了包含图像特征有先验知识的噪声。生成器可以从分布特性明确的噪声中快速搜索到图像的属性特征,解码出噪声先验,同时不低于维度下界的噪声映射到合理的图像特征空间,生成基本符合属性和类别特征的图像。

## 1.2 多阶段图像生成器

### 1.2.1 全局生成器

如图 5(a)所示,随机噪声学习到图像中实例的属性,得到有先验知识的噪声;同时,全局生成器  $G_{img_0}$  计算  $128 \times 128$  语义标签  $L_0$  的全局嵌入向量  $G' \in \mathbb{R}^{D_{\text{emb}}}$ , 结合  $G'$  和获得的噪声先验  $Z_{np}$  进行语义编码,并将二进制实例语义编码聚合为标签映射  $M_i \in \{0, 1\}^{H \times W \times L}$ , 其中  $i \in (1, 2, \dots, T)$  表示实例数,  $W, H$  和  $L$  分别为实例的宽、高和类别标签,当且仅当存在类别为  $k$  且覆盖像素  $(i, j)$  的实例掩码时,即:  $M_{i,j,k} = 1$  时,在该位置进行图像像素表示。计算  $L_0$  的全局嵌入向量  $G'$  的同时,对语义标签  $L_0$  进行下采样得到  $\mu_0$ , 连接  $M_i$  和  $\mu_0$ , 输入到残差块和上采样层,由隐藏层获得图像隐层特征  $y_0$ , 输送到一个  $3 \times 3$  卷积层,初步合成低分辨率的全局图像  $I_0$ , 如式(8)、(9)所示:

$$y_0 = F_0(Z_{np}, G', \text{Enc}(L_0)) \quad (8)$$

$$I_0 = G_{img_0}(y_0) \quad (9)$$

其中:  $\text{Enc}(L_0)$  是低分辨率实例的编码,  $F_0$  被建模为神经网络,  $y_0$  是获得的图像隐层特征。式(9)表示全局生成器  $G_{img_0}$  根据该隐层特征  $y_0$  生成低分辨率图像  $I_0$ 。

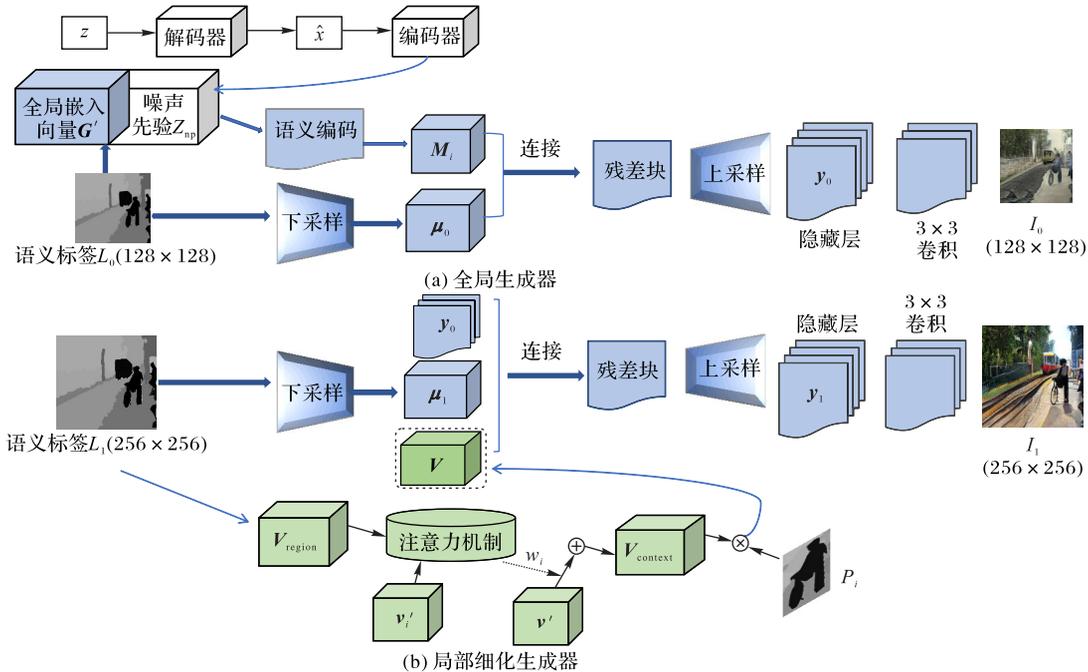


图 5 SLNP-GAN 图像生成器架构

Fig. 5 Architecture of image generators in SLNP-GAN

### 1.2.2 局部细化生成器

全局图像  $I_0$  的生成仅利用全局信息,缺少细粒度实例级别的信息,出现过度平滑的纹理,没有足够的细节和高层次的抽象特征。由于传统的网格注意力机制已成功用于图像-图

像翻译<sup>[16]</sup>和图像问答<sup>[17]</sup>, AttnGAN 将注意力机制引入文本-图像生成任务中,允许简单图像的生成。受此启发,本文在语义标签-图像生成过程中首次引入注意力机制,如图 5(b)所示。局部细化生成器用初步生成的图像子区域向量查询高分辨率

语义标签中的相关实例,获得基于背景信息的实例向量,优化调整以合成匹配实例标签的更准确、细粒度图像。

注意力机制主要有两个方面:首先,根据所有输入信息获得注意力分布;然后,根据注意力分布来计算输入信息的加权平均。将输入信息向量  $X$  作为信息存储器,  $q$  为作为查询向量来选择  $X$  中的相关信息,该过程需要被选择信息的索引。定义变量  $n$  为被选择信息的索引,注意力分布  $\alpha_i$  表示  $X$  中被选择的第  $i$  个信息与查询  $q$  的相关程度。则注意力分布  $\alpha_i$  构成的概率向量为:

$$\alpha_i = p(n = i | X, q) = \frac{\exp(s(x_i, q))}{\sum_{i=1}^T \exp(s(x_i, q))} \quad (10)$$

其中  $s(x_i, q)$  是注意力打分函数,可用点积模型计算:

$$s(x_i, q) = x_i^T q \quad (11)$$

其中:  $x_i$  是输入的第  $i$  个信息, softmax 将权重归一化,得到符合概率分布区间的注意力分配值,用该权重分布表示不同输入受关注的程度。

最后,利用加权平均对输入信息汇总得到注意力值:

$$\text{attn}(X, q) = \sum_{i=1}^T \alpha_i x_i \quad (12)$$

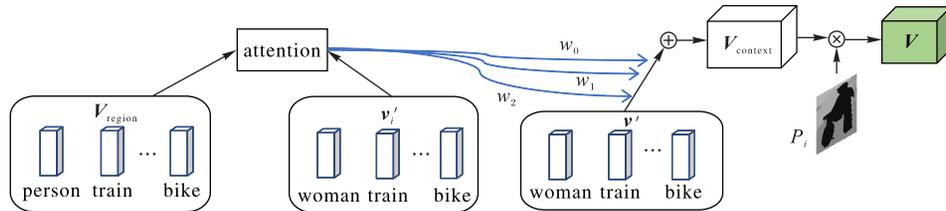


图 6 局部细化生成器的注意力机制

Fig. 6 Attention mechanism of local refined generator

此外,生成过程中语义标签可能有多个像素覆盖同一个像素点,可用实例级别的背景向量解决。为生成外观清晰且真实的图像,必须决定用哪个背景向量对重叠部分像素表示。因此,求解第  $i$  个实例的每个像素  $P_i$  与实例级别的背景向量  $V_{\text{context}}^i$  的向量外积,如式(15)所示:

$$V = \max_{1 \leq i \leq T} P_i \otimes V_{\text{context}}^i \quad (15)$$

其中:  $\otimes$  是向量外积,若多个像素覆盖同一个像素点,则对多个像素点最大池化,使像素  $P_i$  与最相关的实例级别的背景向量  $V_{\text{context}}^i$  关联,在该位置进行像素表示,获取包含底层细节信息向量  $V$ 。

与全局生成器不同,如图 5(b)所示,局部细化生成器为将全局信息从  $G_{\text{img}_0}$  整合到  $G_{\text{img}_1}$ ,  $G_{\text{img}_1}$  残差块的输入是含底层细节信息的向量  $V$  和语义标签  $L_1$  下采样信息  $\mu_1$ ,以及  $G_{\text{img}_0}$  隐藏层的特征  $y_0$ 。然后经上采样获取  $I_1$  的隐藏层特征  $y_1$ ,并输送到  $3 \times 3$  卷积层,由  $G_{\text{img}_1}$  生成  $256 \times 256$  高分辨率图像  $I_1$ 。如式(16)、(17)所示:

$$y_1 = F_1(y_0 + \text{Enc}(L_1), V, V_{\text{context}}) \quad (16)$$

$$I_1 = G_{\text{img}_1}(y_1) \quad (17)$$

其中:  $\text{Enc}(L_1)$  是高分辨率实例编码,  $V$  表示含底层细节信息的向量,  $V_{\text{context}}$  为实例级别的背景向量,  $F_1$  被建模为神经网络。式(17)表示  $G_{\text{img}_1}$  由隐层特征  $y_1$  生成高分辨率图像  $I_1$ 。

因此, SLNP-GAN 采用多阶段的图像生成策略。全局生

局部细化生成器  $G_{\text{img}_1}$  引入的注意力机制,如图 6 所示。

通过关注  $L_1$  中与  $I_0$  子区域实例向量  $V_{\text{region}}$  对应的最相关子标签,来获取实例像素级别的信息,细化不同区域的像素特征。 $G_{\text{img}_1}$  使用  $I_0$  的子区域向量  $V_{\text{region}}$  来查询语义标签  $L_1$  中有更详细信息的相关实例向量  $v_i'$  (如:实例具体为 woman,而非  $I_0$  中  $V_{\text{region}}$  的 person),为每个的实例向量  $v_i'$  分配注意力权重  $w_i$ ,然后由  $w_i$  计算输入信息的加权和,生成基于背景信息的实例向量  $V_{\text{context}}$ ,计算生成图像的第  $j$  个子区域时的背景向量,如式(13):

$$V_{\text{context}}^j = \sum_{i=1}^T w_{j,i} v_i' \quad (13)$$

其中:  $v_i'$  是包含详细信息的第  $i$  个实例向量,生成第  $j$  个子区域时,对第  $i$  个实例分配的注意力权重  $w_{j,i}$  使用注意力机制中权重分布的计算公式求解,得到符合概率分布区间的注意力分配值,如式(14)所示:

$$w_{j,i} = \frac{\exp(s_{j,i})}{\sum_{i=1}^T \exp(s_{j,i})}; s_{j,i} = (V_{\text{region}}^j)^T v_i' \quad (14)$$

其中:注意力打分函数  $s_{j,i}$  采用点积模型计算,使用 softmax 进行权重归一化。

成器结合噪声先验,直接输入语义标签,生成了布局和语义标签基本一致的全局图像;然后局部细化生成器使用注意力机制完善局部细节,生成了  $256 \times 256$  图像。

### 1.3 多级图像判别器与损失函数

为区分真实图像和合成图像,判别器要有较大的感受野,需要更深的网络或更大的卷积内核,会导致容量增加、过拟合和重复图案。为解决该问题,对不同分辨率图像使用相同架构的多级判别器  $D_{\text{img}_0}$  和  $D_{\text{img}_1}$  分别进行训练。判别器架构如图 7 所示。

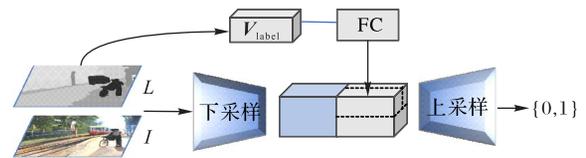


图 7 图像判别器架构

Fig. 7 Architecture of image discriminator

首先,连接生成的图像  $I$  和语义标签  $L$ ,输送到下采样块以产生大小为  $h' \times w'$  的特征映射。同时将  $L$  的标签嵌入向量  $V_{\text{label}}$  全连接并进行空间平铺,经上采样计算判别器的决策分数。虽然二者有相同的结构,但是  $D_{\text{img}_0}$  在粗粒度级别指导  $G_{\text{img}_0}$  生成和语义标签的布局大体一致的图像  $I_0$ ,具有最大的感受野和图像的全局视图;在细粒度级别的  $D_{\text{img}_1}$  用于引导

$G_{img_k}$  生成纹理逼真的  $I_1$ 。将低分辨率模型扩展到高分辨率仅需在细粒度级别添加判别器,无需从头重新训练,由此也使得生成器由粗粒度到细粒度的训练更容易。

图像生成器  $G = \{G_{img_0}, G_{img_1}\}$  和多级判别器的对抗训练是多任务学习过程。GAN 交叉熵损失函数为:

$$\min_G \max_{D_{img_k}} \sum_{k=0,1} \mathcal{L}_{GAN}(G, D_{img_k}) = \mathbb{E}_{x \sim p_{\text{data}}(x)} [\text{lb} D(x)] + \mathbb{E}_{L \sim p_L(L)} [\text{lb}(1 - D(G(L)))] \quad (18)$$

其中:  $x$  表示真实图像,  $D(x)$  表示对真实样本进行判别,判别结果越接近 1,说明模型性能越好;同样,对生成样本  $G(L)$  的判别值越接近 0,也说明模型性能越好。

生成对抗网络训练过程具有不稳定性,易导致模式崩溃,Johnson 等<sup>[18]</sup>提出了感知损失,该损失能对超分辨率图像重构并进行风格转换<sup>[19]</sup>,因此,SLNP-GAN 模型采用与之相关的特征匹配损失<sup>[20]</sup>,从判别器网络的多个层中进行特征提取,比较真实和生成图像的特征,学习匹配真实和合成图像的中间表示,使得生成结果和真实图像接近。将  $D_{img_k}$  的第  $i$  层特征提取器表示为  $D_{img_k}^{(i)}$ ,则特征匹配损失为:

$$\mathcal{L}_{FM}(G, D_{img_k}) = \mathbb{E}_{(L,x)} \sum_{i=1}^T \frac{1}{N_i} \left[ \left\| D_{img_k}^{(i)}(L, x) - D_{img_k}^{(i)}(L, G(L)) \right\|_1 \right] \quad (19)$$

其中:  $N_i$  为每层元素的数量,  $T$  表示总层数,  $L$  和  $X$  分别表示语义标签图和相对应的真实图。

因此,SLNP-GAN 的完整目标损失函数为 GAN 损失函数和特征匹配损失函数的加和,如式(20)所示:

$$\min \left( \max_{D_{img_k}} \sum_{k=0,1} \mathcal{L}_{GAN}(G, D_{img_k}) + \lambda \sum_{k=0,1} \mathcal{L}_{FM}(G, D_{img_k}) \right) \quad (20)$$

其中:  $\mathcal{L}_{GAN}$  为 GAN 损失函数项,  $\mathcal{L}_{FM}$  为特征匹配损失项,  $\lambda$  表示特征匹配损失的权重分配值。

#### 1.4 AMM 算法优化训练

对抗网络训练通常使用 Adam 优化算法,仅计算损失函数的一阶梯度,不同的参数需要设置不同的学习率。由于生成对抗网络的目标函数是复杂高维且非凸的随机函数,该算法在训练时不稳定,可能会跳过全局最优解,导致模型难以收敛<sup>[21]</sup>。

如表 1 所示,本文使用基于 Adam 算法和动量思想提出的 AMM 算法,对 Adam 的参数更新进行了改进。初始时 AMM 算法和 Adam 算法的方向均为  $P_t^{\text{Adam}}$ 。其中:  $\delta$  是常数,初始化为  $10^{-8}$ ,  $\hat{m}_t$  和  $\hat{v}_t$  分别表示一阶矩偏差修正和二阶矩偏差修正,由式(21)、(22)所示:

$$\hat{m}_t = m_t / (1 - \beta_1^t); m_t = \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t; \quad (21)$$

$$\hat{v}_t = v_t / (1 - \beta_2^t); v_t = \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2 \quad (22)$$

其中:  $m_t$  和  $v_t$  表示梯度和梯度平方的指数移动平均,即梯度的一阶和二阶原始矩估计;  $\beta_1^t$  和  $\beta_2^t$  分别表示  $\beta_1$  和  $\beta_2$  的  $t$  次方,两者控制指数移动平均  $m_t$  和  $v_t$  的衰减速率;  $g_t = \nabla_{\theta} f_t(\theta)$  表示第  $t$  个损失函数中关于变量  $\theta$  的梯度向量,即  $f_t$  关于  $\theta$  的偏微分。将  $m_t$  和  $v_t$  的初始化为 0 会产生误差,但是这些误差可以通过

修正消除,从而产生无偏估计  $\hat{m}_t$  和  $\hat{v}_t$ 。

表 1 AMM 和 Adam 参数更新对比

Tab. 1 Parameter update comparison between AMM and Adam algorithm

参数更新步骤	AMM 算法	Adam 算法
初始方向	$P_t^{\text{AMM}} = \frac{\hat{m}_t}{\sqrt{\hat{v}_t + \delta}}$	$P_t^{\text{Adam}} = \frac{\hat{m}_t}{\sqrt{\hat{v}_t + \delta}}$
求更新量	$P_t^{\text{AMM}} \leftarrow \alpha \cdot P_{t-1}^{\text{AMM}} + \eta \cdot P_t^{\text{Adam}}$	$\Delta \theta_t \leftarrow -\eta \cdot P_t^{\text{Adam}}$
参数更新	$\theta_t \leftarrow \theta_{t-1} - P_t^{\text{AMM}}$	$\theta_t \leftarrow \theta_{t-1} + \Delta \theta_t$

计算每一步迭代更新量时,Adam 算法仅和初始学习率以及原来方向  $P_t^{\text{Adam}}$  有关,而 AMM 算法每步迭代更新量为其上一个迭代步与 Adam 当前迭代步的加权之和,其中上个迭代步所占权重为  $\alpha$ ,该更新过程体现了经典动量的思想。最后对时间步为  $t$  的参数  $\theta_t$  进行参数更新时,Adam 算法中,  $\theta_t$  是  $t-1$  时的参数  $\theta_{t-1}$  与更新量  $\Delta \theta_t$  之和,AMM 算法将参数  $\theta_{t-1}$  和第  $t$  步的迭代更新量之差进行参数更新。

相比 Adam 算法,AMM 算法结合了动量和基于  $L_2$  范数优化算法的优点,更加稳定而且收敛速度更快,因此采用 AMM 算法对图像生成任务优化训练,稳定训练过程并加快收敛速度。

## 2 实验结果与分析

### 2.1 实验环境与数据集

本文模型采用深度学习框架 PyTorch 1.2.0,实验环境为 Linux 4.4.0-135-generic 操作系统。使用单个显存为 16 GB 的 Tesla P100 在 COCO\_Stuff 和 ADE20K 数据集上分别训练约 171 h 和 163 h。对于所有数据集,生成器和判别器的学习率设置为 0.0001 和 0.0004,使用 AMM 算法优化训练,一阶矩和二阶矩估计的指数衰减速率  $\beta_1, \beta_2$  设置为 0.9 和 0.999,其中  $\beta_1, \beta_2 \in [0, 1)$ ,常数  $\delta=10^{-8}$ ,初始化历史迭代步所占的权重  $\omega$  和学习率  $\eta$  分别为 0.9 和 0.001。

COCO\_Stuff 数据集<sup>[22]</sup>包含 182 个语义类,具有像素级的标注。按照 COCO\_Stuff 数据集既定的划分,本文使用 118 000 张训练集、5 000 张验证集图像,每张都有 5 句文本描述和对应的语义标签。

ADE20K 数据集<sup>[23]</sup>中的每个文件夹包含对场景分类的图像,对于每一张图像,目标和对象分割被存储为两个不同的文件,所有的图像和对象实例都有注释。该数据集包含 150 个语义类的场景,可用于场景的感知、解析、分割、多物体识别和语义理解。按照 ADE20K 数据集给定的训练集和验证集的划分,实验将 20 210 张图像作为训练集,2 000 张图像作为验证集。对于两个数据集,均使用来自训练集的标签和图像配对数据来训练全局布局和实例像素合成,使用验证集中的语义标签进行图像生成。

### 2.2 实验评价指标

将 SLNP-GAN 模型生成的图像输入语义分割模型,比较预测的语义分割掩码和真实掩码的匹配程度。生成图像与真实图像越相似,则语义分割模型预测到的标签越接近真实标签。采用 DeepLabV3 网络获取平均交并比 (mean Intersection over Union, mIoU) 和像素精准度 (Pixel Accuracy, PA) 指标评估生成图像的准确度:

$$mIoU = \frac{1}{k+1} \sum_{i=0}^k \frac{p_{ii}}{\sum_{j=0}^k p_{ij} + \sum_{j=0}^k p_{ji} - p_{ii}} \quad (23)$$

$$PA = \frac{\sum_{i=0}^k p_{ii}}{\sum_{i=0}^k \sum_{j=0}^k p_{ij}} \quad (24)$$

其中:  $mIoU$  是在每个类的真实值和预测值两个集合的交集和并集之比(Intersection over Union, IoU)的均值,  $k$  为类的个数, 此处定义为  $k+1$  类(包含一个空类或背景);  $p_{ii}$  为真正例数,  $p_{ij}$  和  $p_{ji}$  分别为假负例数和假正例数;  $PA$  表示正确分类的像素占总像素的比例。

同时利用训练好的 Inception v3 网络来提取中间层特征, 用高斯模型的均值  $\mu$  和协方差来计算 Frechet 初始距离(Frechet Inception Distance, FID)。真实图像和生成样本在特征空间的 Frechet 距离表示如式(25)所示:

$$FID(P_r, P_g) = \left\| \mu_r - \mu_g \right\| + T_r(C_r + C_g - 2(C_r C_g)^{1/2}) \quad (25)$$

其中:  $T_r$  表示矩阵对角线上元素的总和,  $C$  是协方差。FID 得分越低, 表示生成的图像与真实图像越接近, 图像质量和多样性更高, 对噪声的鲁棒性更好。

### 2.3 结果分析与对比

SLNP-GAN 进行多阶段图像生成, 如图 8 所示。全局生成器由训练得到的噪声先验学习到粗粒度的属性特征, 并且生成了和语义标签布局一致的全局图像, 但是出现了过度平滑的特征, 缺乏细粒度纹理, 第一阶段的生成结果示例如图

8(a)所示;因此,局部增强生成器结合注意力机制,查询高分辨率语义标签中的相关实例,获得了基于背景信息的实例向量,优化调整合成了匹配实例标签的更准确、细粒度图像,第二阶段的生成结果示例如图 8(b)所示。



图 8 不同阶段生成结果示例

Fig. 8 Result examples of different generation stages

输入相同的语义标签,将 SLNP-GAN 和不同方法的生成结果进行了对比,如图 9 所示。sg2im 方法是由场景图推测语义标签,并使用级联优化网络(Cascaded Refinement Network, CRN)模型将该标签转化为  $64 \times 64$  的图像。由于场景图缺乏核心对象属性和空间交互信息,并且该方法没有引入注意力机制,缺少整体布局的细粒度编码信息,难以在正确的位置生成与与布局一致的相关实体(如图 9(c)踢足球者缺失人体特征)。此外,由于场景结构仅定义了实体和简单的方位信息,未能解决空间位置接近的对象像素重叠问题,无法协调和其他对象的像素表示,导致了难以分离的不同对象外观(如图 9(c)中的公交车未能和背景像素信息区分,不同大象的躯干和轮廓不清晰且出现像素遮挡)。

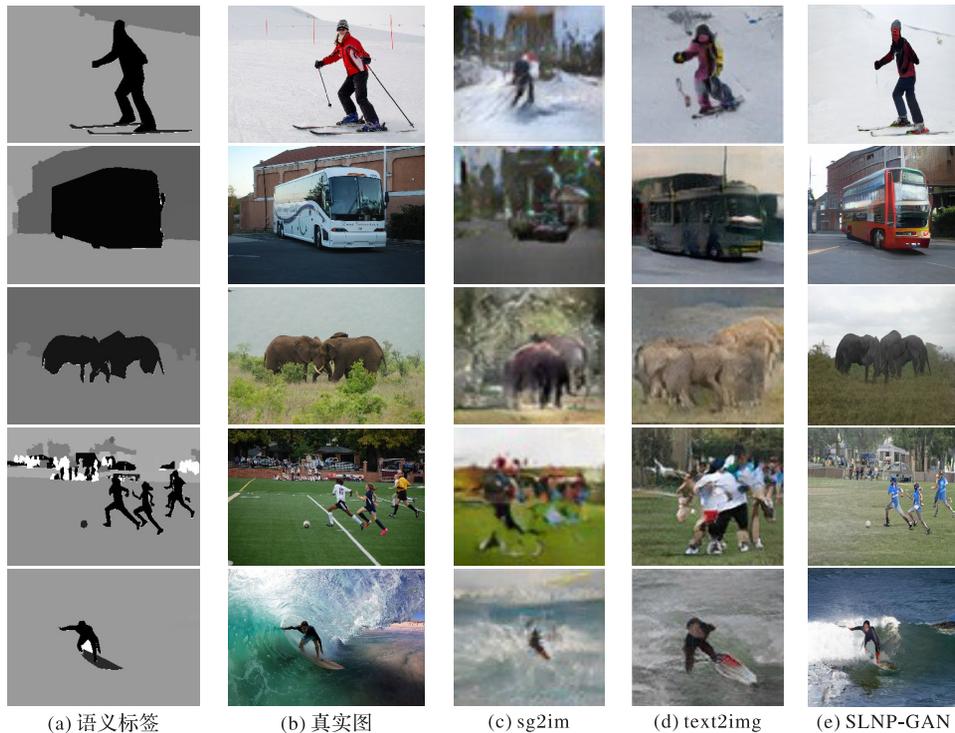


图 9 在 COCO\_Staff 数据集上不同方法生成图像对比

Fig. 9 Comparison of images generated by different methods on COCO\_Staff dataset

text2img 模型输入的是文本,由边界框生成器和形状生成器构建语义标签,最后经图像生成器生成  $128 \times 128$  的图像。该方法是跨文本-图像的多模态生成,由于输入文本中的每个单词都具有描述图像内容的不同级别的信息,但是 text2img

仅以单个句子向量为条件未引入注意力机制,所有实例的权重都相同,没有考虑每个单词对生成结果的影响,缺少每个实例和生成图像整体之间的交互。对于单一对象的生成效果较好(如图 9(d)滑雪者和公交车),但是难以生成包含较多实例

的高分辨率场景,也出现了不同实例难以分离,像素重叠与特征融合的现象(如图9(d)大象和踢足球者)。

而 SLNP-GAN 直接输入语义标签而非将其作为中间表示生成图像,提供了实例位置形状等约束,包含不同实例之间的空间交互关系,因此,在相应的位置都合成了对应的实例布局。同时,加入的噪声先验习得了实例的全局属性,根据布局合成了基本符合现实属性信息的实例。此外,对于不同实例像素重叠的问题,采用对多个像素最大池化,同时结合注意力机制来获得最相关的实例向量,在该位置进行像素表示,解决了不同实例像素遮挡的问题,生成了包含细粒度的纹理特征。另外,相较于其他直接合成图像的方法,由于 SLNP-GAN 采用多阶段生成策略,合成了 256×256 的较高分辨率复杂图像(如图9(e))。

同时,为避免单一数据集可能出现的偏差,使用 ADE20K 数据集也进行实验,如图 10 所示。输入语义标签和噪声先验,SLNP-GAN 经多阶段同样生成了高质量的 256 × 256 图像。由于输入的语义标签提供了全局布局约束,即使对于复杂场景,SLNP-GAN 也能较好地生成符合语义标签的布局。同时,先验知识的噪声作为输入,摒弃了与现实违背的噪声,因此生成的图像几乎没有不合理的属性特征。另外,结合了注意力机制,对于包含较多实例的复杂场景,该模型也能根据权重分配,获取最相关实例向量并进行像素表示,几乎未出现其他模型常见的多个实例难以区分、像素重叠等现象。

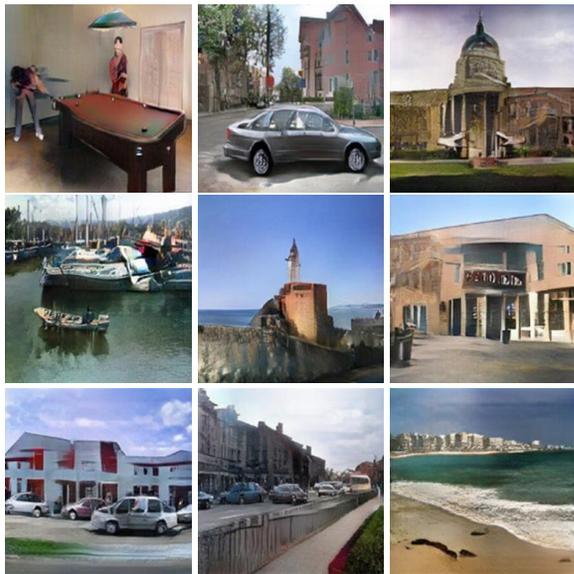


图 10 SLNP-GAN 基于 ADE20K 数据集的生成结果

Fig. 10 Images generated by SLNP-GAN on ADE20K dataset

此外,对噪声先验的效果进行了实验对比,如图 11 所示。SLNP-GAN 在图像生成中加入有先验知识的噪声而非随机噪声。没有加入噪声先验,由于随机噪声包含许多与现实相违背的噪声数据,输入的噪声包含违背现实的特征信息,生成器作为映射函数,只能合成粗粒度的图像,如图 11(c)所示。由于缺少基本的噪声先验约束,随机噪声各个维度随机取值,各向同性,没有侧重性,难以提取到有效特征信息。导致相邻像素之间出现一致的特征,生成的图像整体趋近于单模态且纹理不清晰。加入噪声先验,如图 11(d),模型舍弃了与现实违背的噪声,为噪声增加了先验知识,引导生成器从先验噪声各个维度获取相应的特征信息,从而学习到全局属性和多模态的细节特征,生成了符合真实标签的特征图像。

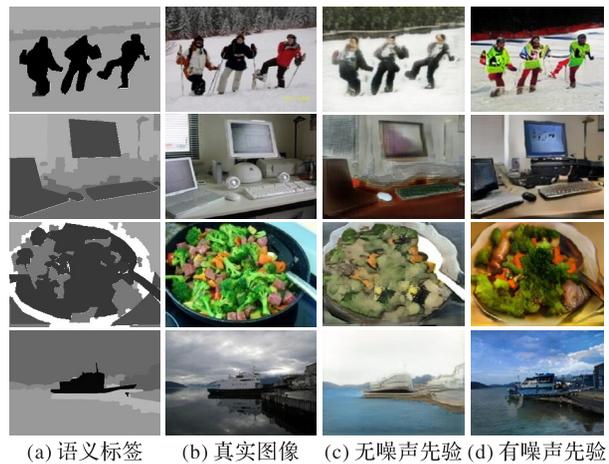


图 11 噪声先验的效果对比

Fig. 11 Effect comparison of noise prior

基于 COCO-Stuff 和 ADE20K 数据集的注意力可视化分别如图 12(a)、(b)所示。图中高亮部分表示生成过程中每一步关注的图像实例区域,在局部细化生成器生成图像的过程中,引入的注意力机制关注生成图像的不同区域并分配注意力,获取语义标签每个位置对应的最相关实例信息,在该位置进行像素级别的图像生成,完善不同实例的细节特征。



图 12 注意力机制可视化图

Fig. 12 Visualization of attention mechanism

为比较 Adam 和 AMM 算法的性能,同时避免单个数据集的误差,在 COCO\_Stuff 和 ADE20K 数据集上均进行了实验。对比结果如图 13、14 所示。

图 13 是 Adam 和 AMM 优化算法在 COCO\_Stuff 数据集上训练时损失值和收敛的变化,图 14 是两算法在 ADE20K 数据集的性能对比。实验均选取了相同的样本量,参数  $\beta_1, \beta_2$  都被初始化为 0.9 和 0.999,学习率均为 0.001,图 13、14 中的  $D_{real}$  和  $D_{fake}$  指标分别代表判别器把生成的图像判为真

和假。刚开始迭代时,二者的损失值在所有数据集上都相近。随着次数的增加,接近 100 000 次时,图 13(a)和图 14(a)中 Adam 的  $D_{fake}$  和  $D_{real}$  对应的损失函数值接近于 0.8,并持续在 0.8 附近波动且幅度较大;而图 13(b)和图 14(b)中 AMM 算法在接近 50 000 次时, $D_{fake}$  和  $D_{real}$  对应的损失值均已趋近于 0.5,并持续在附近波动。对比可知:在不同的数据集上,AMM 算法均能将训练的收敛速度提升大约一倍,并缩短收敛时间;而且在相同的迭代次数条件下,AMM 算法的损失函数值均小于 Adam 算法的损失值,波动幅度更小而且训练更稳定。

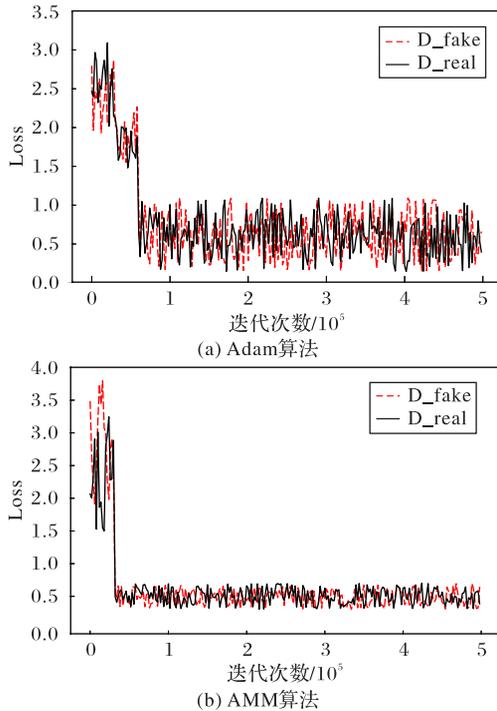


图 13 基于 COCO\_Stuff 数据集的训练性能对比  
Fig. 13 Comparison of training performance on COCO\_Stuff dataset

最后,将生成的图像输入到语义分割网络 DeepLabV3 中得到 mIoU 和 PA 评估值,并用 Inception v3 网络提取中间层特征获得 FID 指标,在 COCO\_Stuff 和 ADE20K 数据集上不同方法的评价指标对比,分别如表 2 和表 3 所示。

表 2 COCO\_Stuff 数据集上不同方法的评价指标对比  
Tab. 2 Comparison of evaluation metrics between different methods on COCO\_Stuff dataset

方法	mIoU	PA	FID
sg2im	21.3	38.6	81.1
text2img	29.7	55.2	41.6
SLNP-GAN	<b>35.1</b>	<b>68.3</b>	<b>33.4</b>

表 3 ADE20K 数据集上不同方法的评价指标对比  
Tab. 3 Comparison of evaluation metrics between different methods on ADE20K dataset

方法	mIoU	PA	FID
sg2im	22.4	39.4	81.8
text2img	32.7	56.8	40.9
SLNP-GAN	<b>38.5</b>	<b>63.1</b>	<b>32.8</b>

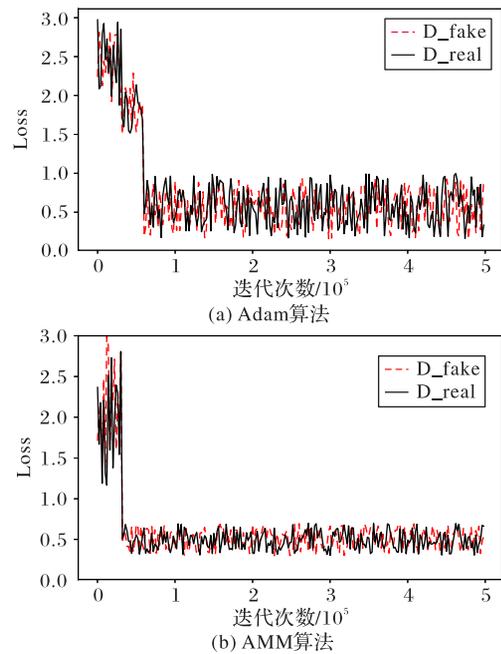


图 14 基于 ADE20K 数据集的训练性能对比  
Fig. 14 Comparison of training performance on ADE20K dataset

表 2、表 3 的结果表明,相较于 text2img,SLNP-GAN 模型在 COCO\_Stuff 和 ADE20K 数据集分别将 mIoU 值提高了 18.18% 和 17.74%,像素准确度 (PA) 增长了 23.73% 和 11.09%,FID 值降低了 19.71% 和 19.80%。由于 sg2im 和 text2img 的输入分别为场景结构和文本,而 SLNP-GAN 直接输入语义标签,能够合成更符合语义布局的图像。因此,将生成结果喂入 Inception v3 网络,得到的语义分割图与真实标签的匹配度较高,得到了较高的平均交并比 (mIoU) 值。此外,其他方法直接输入随机噪声,没有摒弃与现实不符的噪声数据,而本文使用噪声先验,学习全局图像属性特征,生成的图像包含合理的像素特征,像素精准度 (PA) 最高。同时,引入的注意力机制给不同的实例分配了不同的权重,在有像素重叠的区域选择最相关的实例进行像素特征表示,生成结果与真实图像距离较近,FID 最低。而 sg2im 和 text2img 没有区分生成图像的不同实例权重,出现大量的实例遮挡、像素重叠等问题,像素精准度较低,而且生成样本和真实图像在特征空间的距离相差较大 (FID 值较高)。对比可知,SLNP-GAN 使用语义标签直接作为输入,加入噪声先验并结合注意力机制能生成高质量的准确图像。

### 3 结语

针对复杂语义标签生成以实例为中心的图像分辨率不高而且训练效率低的问题,使用基于语义标签和噪声先验的 SLNP-GAN 模型在 COCO\_Stuff 和 ADE20K 数据集上进行真实且高分辨率图像的生成。首先,使用训练获得的噪声先验学习到全局图像属性提升生成结果的准确度,同时用语义标签替代文本或场景图直接作为输入;然后,结合注意力机制生成包含细粒度纹理信息的图像;最后,使用 AMM 算法对图像生成模型进行优化,使得训练更稳定且收敛更快。实验结果表明,SLNP-GAN 模型在不同的数据集上都可以生成分辨率更高的图像、训练过程较稳定而且损失函数值更小。然而图像

生成效率和分辨率仍需进一步提升和完善,后续工作重点将集中于由知识图谱推理得到相应的语义标签,以端到端的方式由一张语义标签生成多张图像以及视频合成的研究。

#### 参考文献(References)

- [1] ZHANG H, XU T, LI H. StackGAN: text to photo-realistic image synthesis with stacked generative adversarial networks [C]// Proceedings of 2017 IEEE International Conference on Computer Vision. Piscataway: IEEE, 2017: 5908-5916.
- [2] REED S, VAN DEN OORD A, KALCHBRENNER K, et al. Parallel multiscale autoregressive density estimation [C]// Proceedings of the 34th International Conference on Machine Learning. New York: JMLR. org, 2017: 2912-2921.
- [3] MANSIMOV E, PARISOTTO E, BA J L. Generating images from captions with attention [EB/OL]. [2019-07-28]. <https://arxiv.org/pdf/1511.02793>.
- [4] CHEN Q, KOLTUN V. Photographic image synthesis with cascaded refinement networks [C]// Proceedings of the 2017 IEEE International Conference on Computer Vision. Piscataway: IEEE, 2017: 1520-1529.
- [5] ISOLA P, ZHU J, ZHOU T, et al. Image-to-image translation with conditional adversarial networks [C]// Proceedings of the 2017 IEEE International Conference on Computer Vision and Pattern Recognition, 2017: 5967-5976.
- [6] QIAO T, ZHANG J, XU D, et al. MirrorGAN: learning text-to-image generation by redescription [C]// Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2019: 1505-1514.
- [7] HONG S, YANG D, CHOI J, et al. Inferring semantic layout for hierarchical text-to-image synthesis [C]// Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 7986-7994.
- [8] JOHNSON J, GUPTA A, LI F. Image generation from scene graphs [C]// Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 1219-1228.
- [9] LI Y, OUYANG W, ZHOU B, et al. Scene graph generation from objects, phrases and region captions [C]// Proceedings of 2017 IEEE International Conference on Computer Vision. Piscataway: IEEE, 2017: 1270-1279.
- [10] 郭丽丽, 于世飞. 深度学习研究进展 [J]. 计算机科学, 2015, 42(5): 28-33. (GUO L L, DING S F. Research progress on deep learning [J]. Computer Science, 2015, 42(5): 28-33.)
- [11] 刘波宁, 翟东海. 基于双鉴别网络的生成对抗网络图像修复方法 [J]. 计算机应用, 2018, 38(12): 3557-3562, 3595. (LIU B N, ZHAI D H. Image completion method of generative adversarial networks based on two discrimination networks [J]. Journal of Computer Applications, 2018, 38(12): 3557-3562, 3595.)
- [12] XU T, ZHANG P, HUANG Q, et al. AttnGAN: fine-grained text to image generation with attentional generative adversarial networks [C]// Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 1316-1324.
- [13] MNH V, HEES N, GRAVES A, et al. Recurrent models of visual attention [C]// Proceedings of the 27th International Conference on Neural Information Processing Systems. Cambridge: MIT Press, 2014: 2204-2212.
- [14] WANG T, LIU M, ZHU J, et al. High-resolution image synthesis and semantic manipulation with conditional GANs [C]// Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 8798-8807.
- [15] ANDERSON P, HE X, BUEHLER C, et al. Bottom-up and top-down attention for image captioning and visual question answering [C]// Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 6077-6086.
- [16] MA S, FU J, CHEN C W, et al. DA-GAN: instance-level image translation by deep attention generative adversarial networks [C]// Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 5657-5666.
- [17] LU J, YANG J, BATRA D, et al. Hierarchical question-image co-attention for visual question answering [EB/OL]. [2019-06-23]. <https://arxiv.org/pdf/1606.00061.pdf>.
- [18] JOHNSON J, ALAHI A, LI F. Perceptual losses for real-time style transfer and super-resolution [C]// Proceedings of 2016 European Conference on Computer Vision, LNCS 9906. Cham: Springer, 2016: 694-711.
- [19] GATYS L A, ECKER A S, BETHGE M. Image style transfer using convolutional neural networks [C]// Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2016: 2414-2423.
- [20] 修春波, 马云菲, 潘肖楠. 基于距离融合的图像特征点匹配方法 [J]. 计算机应用, 2019, 39(11): 3158-3162. (XIU C B, MA Y F, PAN X N. Image feature point matching method based on distance fusion [J]. Journal of Computer Applications, 2019, 39(11): 3158-3162.)
- [21] LI Y, SNAVELY N, HUTTENLOCHER D P. Location recognition using prioritized feature matching [C]// Proceedings of 2010 European Conference on Computer Vision, LNCS 6312. Berlin: Springer, 2010: 791-804.
- [22] CAESAR H, UIJLINGS J, FERRARI V. COCO-stuff: thing and stuff classes in context [C]// Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 1209-1218.
- [23] ZHOU B, ZHAO H, PUIG X, et al. Scene parsing through Ade20K dataset [C]// Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2017: 5122-5130.

This work is partially supported by the Youth Program of National Science Foundation of China (61601261), the Program of Graduate Education Quality Improvement of Shandong Province (SDYY17136), the Interdisciplinary Research Project of Qufu Normal University (QFNUSKC 291809120).

**ZHANG Susu**, born in 1997, M. S. candidate. Her research interests include computer vision, image generation, deep learning.

**NI Jiancheng**, born in 1971, Ph. D., professor. His research interests include computer vision, machine learning, distributed computing.

**ZHOU Zili**, born in 1973, Ph. D., associate professor. His research interests include intelligent information processing, embedded intelligence, knowledge engineering.

**HOU Jie**, born in 1996, M. S. candidate. Her research interests include computer vision, deep learning.